Universidade Federal de Santa Catarina
Centro de Ciências Físicas e Matemáticas
Programa de Pós-Graduação em Matemática Pura e Aplicada

Alek Fröhlich

# Elements of learning theory and their application in the prediction of malignancy of breast lesions

Florianópolis
2024

Alek Fröhlich

# Elements of learning theory and their application in the prediction of malignancy of breast lesions

Dissertação submetida ao Programa de Pós-Graduação em Matemática Pura e Aplicada para a obtenção do título de mestre em Matemática.
Orientador: Prof. Douglas Soares Gonçalves, Dr.
Coorientador: Prof. Daniel Guimarães Tiezzi, Dr.

Florianópolis
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Fröhlich, Alek
Elements of learning theory and their application in the prediction of malignancy of breast lesions / Alek Fröhlich ; orientador, Douglas Soares Gonçalves, coorientador, Daniel Guimarães Tiezzi, 2024.
96 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências Físicas e Matemáticas, Programa de Pós-Graduação em Matemática Pura e Aplicada, Florianópolis, 2024.

Inclui referências.

1. Matemática Pura e Aplicada. 2. Aprendizagem de Máquina. 3. Teoria de Aprendizagem de Máquina. 4. Modelagem de Predição Clínica. 5. Câncer de Mama. I. Soares Gonçalves, Douglas. II. Guimarães Tiezzi, Daniel. III. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Matemática Pura e Aplicada. IV. Título.

Alek Fröhlich

**Elements of learning theory and their application in the prediction of malignancy of breast lesions**

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Roberto Imbuzeiro Moraes Felinto de Oliveira, Dr.
Instituto Nacional de Matemática Pura e Aplicada

Prof. Alexandre Dias Porto Chiavegatto Filho, Dr.
Universidade de São Paulo

Prof. Luiz-Rafael Santos, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Matemática.

ICP-Edu
Documento assinado digitalmente
**Douglas Soares Goncalves**
Data: 17/06/2024 17:02:10-0300
CPF: ***.566.318-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Douglas Soares Gonçalves, Dr.
Coordenador do Programa

ICP-Edu
Documento assinado digitalmente
**Douglas Soares Goncalves**
Data: 17/06/2024 17:02:30-0300
CPF: ***.566.318-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Douglas Soares Gonçalves, Dr.
Orientador

Florianópolis, 2024.

*À minha namorada Maria Eduarda.*

# Acknowledgements

I am profoundly grateful to Professor Douglas for his guidance and support during the writing of this thesis. His meticulous revision of my work and steadfast commitment ensured that I could meet my deadlines, despite navigating a significant transition from operator algebras to machine learning in my second year. I am also very grateful to Professor Daniel Tiezzi, who introduced me to a breath of interesting problems in breast cancer research. I am particularly thankful for his warm hospitality during my visit to Ribeirão Preto and the University of São Paulo's Medical School. He has been a good mentor, and more importantly, a good friend all throughout the course of my master's project. I sincerely hope we are able to maintain contact in the future.

I am deeply indebted to Isabela Buzatto, a dedicated physician from the University of São Paulo's Hospital das Clínicas and also my academic sibling. The breast lesion malignancy prediction project used as a case study in this thesis stems from her doctoral research on the same subject. Therefore, she is to be credited for all the networking and data collection efforts upon which the experiments in this thesis were built. Moreover, in her role as one of the ultrasound operators at the hospital, she not only envisioned the project but also served as my primary source of domain expertise for the experiments conducted herein. I firmly believe that such collaborative interactions form the cornerstone of meaningful interdisciplinary research, and I am profoundly grateful for the opportunity to collaborate on this project.

I would like to thank all other professors that contributed to my master's journey. In particular, I would like to thank Professors Alcides and Ruy for overseeing me during my first year, when I still considered pursuing a path in operator algebras. They, together with professors Pestov, Douglas, and Daniel G., were particularly important in guiding me during my transition period. Furthermore, I would like to thank Professor Gilles for all the extra lectures on cool topics, such as the Stone–Čech compactification, that he gave me during his point-set topology course. In a similar spirit, I would like to thank Professor Matheus for his thrilling summer course on functional analysis. Finally, I would also like to thank the IMPA professors that made my stay in Rio more pleasant and which helped me during my PhD application process: Benar F. Svaiter, João Pereira, Lucas Nissenbaum, and Roberto Imbuzeiro.

On a personal note, I would like to thank my girlfriend for her constant love and support. Her passion for working hard truly inspires me and has been a motivating factor for me to push myself harder lately. I would also like to thank my parents, Guto e Lu, for all their support and counsel in academic matters. Furthermore, I would like to thank all the friends I made during my masters: Ricardo, Luíz, Francisco, João, and Bernardo from UFSC, Gustavo from USP, and Christian, Otávio, Igor, Daniel P., Marcos, Eduardo, Leonardo(s), Manoel, Arthur, Yangrui, Zhifei, and Di Liu from IMPA. I would also like to thank my lifelong friends: João P. (Jota), Victor M. (Mosi), and Leonardo D. (Dute).

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

*"As the power of computer approaches the theoretical limit and as we turn to more realistic (and thus more complicated) problems, we face 'the curse of dimension', which stands in the way of successful implementations of numerics in science and engineering. Here one needs a much higher level of mathematical sophistication in computer architechture as well as in computer programming. …Successes here may provide theoretical means for performing computations with high power growing arrays of data. …We shall need … the creation of a new breed of mathematical professionals able to mediate between pure mathematics and applied science. The cross-fertilization of ideas is crucial for the health of the science and mathematics."*

*Mikhaïl Gromov*

*"We must not believe those, who today, with philosophical bearing and deliberative tone, prophesy the fall of culture and accept the ignorabimus. For us there is no ignorabimus, and in my opinion none whatever in natural science. In opposition to the foolish ignorabimus our slogan shall be Wir müssen wissen – wir werden wissen (We must know – we will know)."*

*David Hilbert*

# Resumo

A estratégia atual de controle do câncer de mama no sistema público de saúde brasileiro depende da determinação manual de escores BI-RADS para avaliação de malignidade durante exames de ultrassom, frequentemente resultando em biópsias desnecessárias. A previsão de malignidade a partir de características clínicas e de ultrassom poderia aliviar a carga de trabalho dos patologistas e compensar lacunas de habilidade em médicos iniciantes ou não especialistas. Métodos de aprendizagem de máquina têm se mostrado promissores no uso de características de ultrassom de modo B para previsão de malignidade de lesões de mama. Nesta dissertação, discutimos elementos da teoria de aprendizagem de máquina, incluindo desigualdades de concentração e dimensão VC, que são conceitos-chave para a análise de propriedades de generalização de algoritmos de aprendizagem. Em seguida, mostramos como tais conceitos podem ser usados para elaboração de cotas de generalização para os valores preditivos. Em cenários com grandes tamanhos de amostra e pequena dimensão VC, um estudo de validação baseado nessas cotas de generalização seria possível. Também apresentamos uma abordagem baseada em gradient boosting para identificação de lesões benignas, que incorpora características clínicas, baseadas em Doppler e clássicas de ultrassom em modo B. Um classificador XGBoost foi treinado com dados de 1929 lesões de mama obtidas a partir de uma coorte de pacientes de quatro centros de referência de câncer de mama no Brasil. Nosso classificador alcançou uma área sob a curva de precisão-recall média (AUPRC) de 0,95 e boa calibração em validação cruzada repetida de 5 folds. Nosso trabalho fornece uma solução baseada em gradient boosting promissora que pode beneficiar a prática clínica. Embora não seja aplicável para estimar o erro de generalização das curvas de valor preditivo em nosso problema, devido a um tamanho de amostra insuficiente e à falta de precisão nas estimativas atuais para a dimensão VC de gradient boosted regression trees, as ferramentas matemáticas desenvolvidas nesta dissertação são de extrema importância para o design de algoritmos de aprendizagem confiáveis e podem ser aplicadas a uma gama mais ampla de problemas do que os considerados nesta dissertação.

**Palavras-chave:** Aprendizagem de Máquina. Teoria de Aprendizagem de Máquina. Modelagem de Predição Clínica. Câncer de Mama. Ultrassom.

# Abstract

The current breast cancer control strategy employed in the Brazilian public health system relies on manual determination of BI-RADS scores by clinicians to assess malignancy during ultrasound examinations, often leading to unnecessary biopsies. Predicting malignancy from clinical and ultrasound features could ease pathologists' workload and offset skill gaps in beginner or non-specialist physicians. Machine learning has shown promise in using B-mode ultrasound features to predict breast lesion malignancy. In this thesis, we discuss elements from the theory of machine learning, including concentration inequalities and VC dimension, which are key concepts to analyse the generalization properties of learning algorithms. Then, we show how such concepts can be used to devise generalization bounds for the predictive values. In scenarios with large sample sizes and small VC dimension, a learning-theoretical validation study based on these predictive value generalization bounds would be possible. We also present a gradient boosting approach for identifying benign lesions that incorporates both clinical and Doppler-based features alongside classical B-mode ultrasound features. An XGBoost classifier was trained with data from 1929 breast lesions obtained from a cohort of patients across four breast cancer reference centers in Brazil. Our classifier achieved a mean area under the precision-recall curve (AUPRC) of 0.95 and good calibration in repeated 5-fold cross-validation. Our work provides a promissing gradient boosting solution that may be of benefit to clinical practice. Although not applicable to the estimation of the generalization error of the predictive value curves in our problem due to an insufficient sample size and lack of tightness in current estimates of the VC dimension of gradient-boosted trees, the mathematical tools developed in this thesis are of utmost importance to the design of reliable learning algorithms and may be applied to a wider range of problems than the ones considered in this thesis.

**Keywords:** Machine Learning. Learning Theory. Clinical Prediction Modeling. Breast Cancer. Ultrasound.

# Resumo Estendido

## Introdução

A estratégia atual de controle do câncer de mama no sistema público de saúde brasileiro depende da determinação manual de escores BI-RADS para avaliação de malignidade durante exames de ultrassom, frequentemente resultando em biópsias desnecessárias. Neste cenário, é evidente que a previsão precisa de malignidade a partir de características clínicas e de ultrassonografia poderia aliviar a carga de trabalho dos patologistas e compensar lacunas de habilidade em médicos iniciantes ou não especialistas. Entre as abordagens computacionais para o auxílio à interpretação de exames de ultrassom, aquelas baseadas em aprendizagem de máquina têm se mostrado particularmente promissoras. De um ponto de vista matemático, tais metodologias podem ser analisadas no contexto de teoria do aprendizado, uma área fértil na interseção da matemática aplicada, computação e estatística, que visa entender, entre outras coisas, as propriedades de generalização de algoritmos de aprendizagem de máquina. Alguns elementos relevantes incluem as desigualdades de concentração, as noções de complexidade para classes de modelos e o princípio/algoritmo da minimização do risco empírico. Através desse ferramental, torna-se possível estudar propriedades de generalização de métricas relevantes no contexto diagnóstico, em particular, os valores preditivos positivo e negativo.

## Objetivos

Diante do exposto, esta dissertação de mestrado tem os seguintes objetivos:

1. Estudar e apresentar elementos da teoria do aprendizado supervisionado;
2. Analisar o algoritmo da minimização do risco empírico;
3. Apresentar cotas de generalização para os valores preditivos;
4. Desenvolver e validar um modelo de aprendizagem de máquina para predição de malignidade.

## Metodologia

Nesta dissertação, discutimos elementos da teoria de aprendizagem de máquina, incluindo desigualdades de concentração, complexidade de Rademacher e dimensão VC, que são conceitos-chave para a análise de propriedades de generalização de algoritmos de aprendizagem. Em seguida, mostramos como tais conceitos podem ser usados para elaboração de cotas de generalização para os valores preditivos, métricas centrais no contexto diagnóstico. Também apresentamos experimentos baseados em gradient boosting para identificação de lesões benignas, que incorporam características clínicas, baseadas em Doppler e clássicas de ultrassom em modo B. Em particular, treinamos um classificador XGBoost com base em validação cruzada repetida de 5 folds utilizando dados de 1929 lesões de mama obtidas a partir de uma coorte de pacientes de quatro centros de referência de câncer de mama no Brasil, incluindo o Hospital das Clínicas de Ribeirão Preto e o Hospital de Amor de Barretos. Adicionalmente, apresentamos um segundo conjunto de experimentos que visou avaliar num contexto prático as cotas de generalização para os valores preditivos.

# Resultados e Discussão

Nosso modelo de classificação alcançou boas curvas de valores preditivos, ROC, de precisão e recall, sensitividade, especificidade, e de calibração ao longo das 50 iterações de validação cruzada. Em particular, todas as curvas se mostraram estáveis, as curvas de classificação se mostraram altamente discriminatórias (AUPRC média $= 0.9527$; AUROC média $= 0.9568$) e as curvas de calibração se mantiveram perto do ideal ($y = x$). Com base nos hiperparâmetros escolhidos durante a validação cruzada, desenvolvemos um classificador final, que similarmente atingiu boas propriedades de discriminação e calibração. A fim de inspecionar as predições individuais do modelo, computamos os valores SHAP associados e visualizamos uma importância conjunta de atributos clínicos e de imagem. Por fim, os experimentos baseados nas cotas de generalização não obtiveram resultados interessantes no contexto do nosso dataset e método (XGBoost). Entretanto, notamos que, em cenários com grandes tamanhos de amostra e pequena dimensão VC, um estudo de validação baseado nessas cotas de generalização seria possível.

# Considerações Finais

Os experimentos descritos nesta dissertação possuem vários desdobramentos interessantes. Dentre eles, destacamos a realização de um estudo de validação externo detalhado, estratificando por escore BI-RADS, idade e outros parâmetros importantes. Neste momento, seria interessante avaliar as predições individuais do modelo final com respeito a valores SHAP e também no contexto de predição conforme, com a finalidade de identificar possíveis vieses no modelo e também prover quantificação de incerteza. De um ponto de vista matemático, notamos a importância de se estudar cotas dependentes da distribuição subjacente dos dados e, em particular, baseadas na complexidade de Rademacher. Por fim, embora não sejam aplicáveis para estimar o erro de generalização das curvas de valor preditivo em nosso problema, devido a um tamanho de amostra insuficiente e imprecisão nas estimativas atuais para a dimensão VC de gradient boosted regression trees, as ferramentas matemáticas desenvolvidas nesta dissertação são de extrema importância para o design de algoritmos de aprendizagem confiáveis e podem ser aplicadas a uma gama mais ampla de problemas do que os considerados nesta dissertação.

**Palavras-chave:** Aprendizagem de Máquina. Teoria de Aprendizagem de Máquina. Modelagem de Predição Clínica. Câncer de Mama. Ultrassom.

# List of Figures

# List of Tables

# List of Symbols

| | |
|---|---|
| $\mathcal{X}$ | Feature space |
| $\mathcal{Y}$ | Label space |
| $\mathcal{Z}$ | Joint space $\mathcal{X} \times \mathcal{Y}$ |
| $\Omega$ | Probability space |
| $\mathbb{P}$ | Probability measure / distribution |
| $\mathbb{E}[\cdot]$ | Expected value |
| $\overset{iid}{\sim}$ | Independent and identically distributed |
| $\overset{\mathbb{P}}{\to}$ | Convergence in probability |
| $x$ | Data vector |
| $y$ | Label |
| $\#$ | Cardinality of a set |
| $\mathbf{1}$ | Indicator function |
| $\mathcal{H}$ | Hypothesis class |
| $\mathcal{G}, \mathcal{F}$ | Function class |
| $h$ | Classifier |
| $f$ | Scoring function |
| $\hat{h}_m$ | Empirical risk minizer / approximate empirical risk minimizer |
| $R(h)$ | Risk of classifier $h$ |
| $\hat{R}_m(h)$ | Empirical risk of classifier $h$ with respect to a sample of size $m$ |
| $\hat{\mathrm{err}}_m(\mathcal{H})$ | $\sup_{h\in\mathcal{H}} \lvert R(h) - \hat{R}_m(h)\rvert$ |
| $\mathrm{err}_m(\mathcal{H})$ | Expected value of $\hat{\mathrm{err}}_m(\mathcal{H})$ |
| $\hat{\mathfrak{R}}_m(\mathcal{G})$ | Empirical Rademacher complexity of $\mathcal{G}$ |
| $\mathfrak{R}_m(\mathcal{G})$ | Rademacher complexity of $\mathcal{G}$ |
| $\Pi(\mathcal{G}, m)$ | Growth function of $\mathcal{G}$ with respect to a sample of size $m$ |
| $\Theta(\mathcal{F}, m, k)$ | $(m, k)$-order coefficient |
| VC-dim | Vapnik-Chervonenkis dimension |
| VC-sub | Vapnik-Chervonenkis subgraph dimension |
| $\mathrm{p\hat{p}v}$ | Empirical positive predictive value |
| $\mathrm{n\hat{p}v}$ | Empirical negative predictive value |
| ppv | Positive predictive value |
| npv | Negative predictive value |
| $\hat{q}_f(\cdot)$ | Empirical quantile function |
| $q_f(\cdot)$ | Quantile function |

# Contents

# 1 Introduction

> There were few successes in the treatment of disseminated cancer. It was usually a matter of watching the tumor get bigger, and the patient, progressively smaller.
>
> *John Laszlo*

Machine learning is quickly gaining ground within the medical community as a tool for enabling automation and the discovery of new scientific facts [Deo15; Pic+21; RDK19; Tie+23b]. In this thesis we explore the usage of gradient-boosted regression trees to develop a model for the prediction of malignancy of breast lesions identified by ultrasound in Brazil. A key concern of the thesis is to develop a model that reliably rejects unnecessary biopsies without compromising cancer cases, thereby improving on the current BI-RADS[1] standard.

At its core, machine learning is concerned with leveraging regularities present in natural processes to obtain patterns that *generalize*; i.e., that continue to hold in the presence of new data. In the context of the malignancy prediction problem, this means to find some function $h$ of the observed attributes that reliably predicts the malignancy of breast lesions not present in its training dataset. Naturally, there is much interest in developing mathematical models of learning so that one can analyze which factors most strongly affect it, develop new algorithms, and provide statistical guarantees on practical experiments. For these reasons, in this thesis we explore the classical framework of *learning theory*[2] [BBL02; BLM16; MRT18; Vap98] and apply it to estimate the generalization error of predictive value curves in our machine learning experiments.

## 1.1 Breast lesion malignancy prediction

Breast cancer is the most common invasive cancer among women [Sie+23]. In 2022 alone, there were an estimated 2.3 million new cases and 666.000 deaths worldwide

[1] BI-RADS is a standard for classifying findings based on their malignancy risk (from 0 up to 6). See [MBB+13] for further details.

[2] What is nowadays referred to as learning theory is the combination of two lines of research: *statistical learning theory* and *computational learning theory*. The first originated with the theoretical analysis of the Perceptron by A. Novikoff, M. Aizerman, E. Braverman, and L. Rozonoer [ABR64; MP43; Nov62; Ros62] and blossomed with the analysis of empirical risk minimization and model complexity by V. Vapnik and A. Chervonenkis [VC68; VC74; VC89]. The second is centered around the introduction of the Probably Approximately Correct (PAC) model of learnability to theoretical computer science by Valiant [Val84]. Although originally developed with different goals, both fields shared many tools and techniques and were eventually considered to be part of the more general framework nowadays referred to as learning theory.

[Bra+24]. In Brazil, the situation is not different [INC23]. Breast cancer screening is widely employed to identify early-stage cancers, allowing for easier and less costly treatment [Pra+18]. While mammography serves as the primary technique, it may not always yield conclusive results, particularly in cases involving dense breasts [Ber08]. In such instances, supplemental breast ultrasound[3] (US) is frequently utilized as an alternative. Additionally, US is the recommended tool for the evaluation of many breast abnormalities [Eva+18].

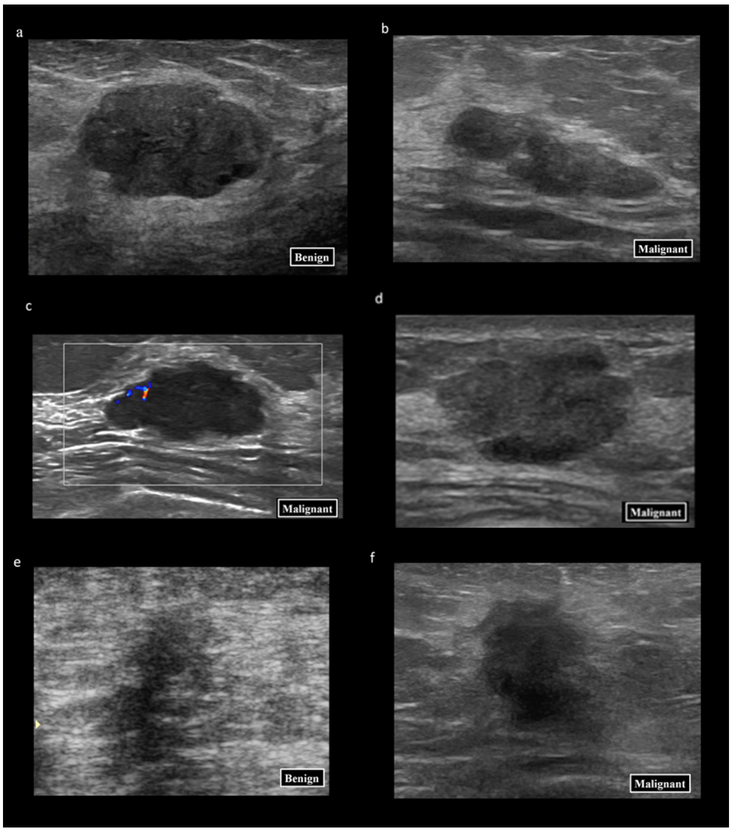


Figure 1 – US samples from benign and malignant lesions.

Source: [Buz+23].

US has several advantages over other imaging methods, including its relatively lower cost, absence of ionizing radiation, and real-time image evaluation capabilities [Fei10]. However, it has a major problem: it is highly operator-dependent and exhibits a high false positive rate [Cor+11; Yan+20]. Moreover, this issue is only partially ameliorated by employing the widely adopted Breast Imaging Reporting & Data System (BI-RADS) standard [Kim+21; SBC20], resulting in a substantial number of unnecessary biopsies, added burden on pathologists, and instilled fear on patients and families [Buz24]. This underscores the need for improved diagnostic strategies to mitigate unnecessary biopsies and enhance the precision of ultrasound-based evaluations.

There have been numerous studies employing the use of machine learning to improve the diagnostic performance of US [Cho+19; Lee+12; Mou+20; Pfo+22; She+07;

[3] Refer to Figure 1 for an example of US imaging.

She+23; She+21; Zel+19]. In this thesis, we extend the work of Buzatto et al. [Buz+23] by exploring a learning-theoretic approach to malignancy prediction and considering alternative model validation techniques based on predictive value generalization bounds [VS20] and repeated cross-validation. Our experiments build on the dataset collected by Buzatto et al. in collaboration with four breast cancer reference centers in Brazil: University of São Paulo's University Hospital at Ribeirão Preto (SP), Women's Health Reference Center of Ribeirão Preto (MATER) (SP), Hospital de Amor de Barretos (SP), and Hospital de Amor de Campo Grande (MS).

Although clinical prediction modeling has a long relationship with statistical inference [Ste19], the machine learning approach taken in this thesis is better discussed within the context of learning theory. There are two reasons for this. The first is that gradient boosting originates from learning theory [CG16; FS97; FHT00; Fri01; KV94]. The second is that the theory enables the estimation of the generalization error of predictive value curves under the relatively mild assumption that the flow of suspicious lesions through the clinic is given by independent and identically distributed (i.i.d.) sampling from a fixed but unknown probability distribution $\mathbb{P}$. From a more general perspective, it is also worth highlighting that most supervised learning approaches, including our own, follow the inductive principle of empirical risk/loss minimization (we minimize a regularized version of the log-loss), and, as such, there is intrinsic motivation for studying it. Moreover, it must be noted that concentration inequalities are a basic tool to the non-asymptotical analysis of many random processes [BLM16], both inside and outside of machine learning, and as such are applicable to a wider range of problems than the ones considered in this thesis.

## 1.2 Goals

The objectives of this dissertation are as follows.

- **Present elements of learning theory:** In this thesis, we review central concepts from learning theory. In particular, we study the concentration inequalities of *Chernoff-Hoeffding* (Theorem 2.3.1) and *McDiarmid* (Theorem 2.3.3), and the notions of *Rademacher complexity* (Definition 2.4.1) and *VC dimension* (Definition 2.4.4).

- **Analyze empirical risk minimization:** Simultaneously with the introduction of the mathematical tools of model complexity and concentration, we study the empirical risk minimization principle as a learning algorithm and show how such tools may be used to analyse its learning rate.

- **Provide a learning-theoretic treatment of malignancy prediction:** After reviewing foundational material, we specialize the discussion to malignancy prediction. In particular, we explore relevant classification metrics, discuss scoring functions and threshold-based classifiers, and provide a learning-theoretic analysis of the generalization error of predictive value curves that builds on VC dimension and concentration.

- **Develop and validate a machine learning model:** Finally, we experiment with the dataset described in [Buz+23] to develop and validate a gradient-boosted trees model for predicting the malignancy of breast lesions identified by ultrasound in the Brazilian public health system.

## 1.3 Structure of the thesis

This dissertation is organized as follows. Chapter 2 introduces concentration inequalities and abstract notions of model complexity wrapped around a discussion on empirical risk minimization. The chapter begins with a thorough introduction to the binary classification problem and the empirical risk minimization algorithm. Following this, the central inequalities of *McDiarmid* and *Chernoff-Hoeffding* are introduced. Then, two formalizations of the intuitive notion of model complexity are discussed: *Rademacher complexity* and *VC dimension*. The chapter culminates with the usage of such complexity measures to provide non-asymptotical generalization bounds for empirical risk minimization over infinite classes. Chapter 3 starts with an introduction to relevant performance metrics. Following this, we discuss threshold-based classifiers and the learning-theoretic generalization bounds of Vemuri and Srebro [VS20]. Chapter 4 revolves around two experiments. The first experiment illustrates the theory of chapters 2 and 3 in the context of gradient boosting and our dataset. Given the high uncertainty present in the performance bounds obtained in the first experiment, the second experiment relies on more empirical methods for hyperparameter tuning and internal validation. Chapter 5 discusses both the theoretical material of Chapters 2 and 3 as well as the experiments of Chapter 4. The thesis ends with a description of possible extensions of this work.

# 2 Mathematical foundations of supervised learning

> I heard reiteration of the following claim: Complex theories do not work; simple algorithms do. I would like to demonstrate that in the area of science a good old principle is valid: Nothing is more practical than a good theory.
>
> *Vladimir Vapnik*

In this chapter, we delve into the mathematical underpinnings of supervised learning. We focus particularly on binary classification, as all applied learning problems of this thesis can be cast into this framework. For multiclass classification and regression, readers can refer to [MRT18, Chapters 9 and 11].

In the context of machine learning, the problem of binary classification consists in learning to discern between two classes denoted by the labels $+1$ and $-1$. Specifically, given a training dataset $\{(x_i, y_i)\}_{i=1}^m$, where $x_i \in \mathcal{X}$ (usually $\mathcal{X} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$) and $y_i \in \{-1, 1\}$, the goal is to construct a classifier function $h : \mathcal{X} \to \{-1, 1\}$ capable of predicting the classes of new, unseen data points; i.e., to construct a classifier that *generalizes* beyond the training dataset. In practice, generalization is assessed by comparing the classifier's accuracy in predicting labels for the training data $\{(x_i, y_i)\}_{i=1}^m$ to its performance on a separate test set $\{(x_i, y_i)\}_{i=m+1}^n$, where $n > m$. In this setting, a classifier demonstrating small error on both datasets is said to have *generalized*, while a classifier with low training error but high test error is deemed *overfitted* to the training data. Furthermore, a classifier exhibiting high error in both training and test sets is said to be *underfitted* to the training data. For a concrete example of binary classification, readers are referred to [SWM93], which discusses the problem of classifying breast fine needle aspirates as malignant or benign.

The construction of the classifier can be approached in multiple ways. For instance, a group of domain experts may handcraft a model[1] whose weights reflect their prior knowledge about the problem. More common, however, is to algorithmically extract a model from the data; i.e., to let the *machine learn*. This can also be done in numerous ways, but a very common strategy is to consider a collection of classifiers $\mathcal{H}$ and to choose $h \in \mathcal{H}$ that best fits the data[2]. Two noteworthy examples of model classes are given by

[1] We will use the words classifier and model interchangeably.

[2] In machine learning, fitness is measured by a loss function $L : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ that quantifies the

the class of linear classifiers, defined as $\mathcal{H}_{\text{LIN}} = \{x \mapsto \text{sign}(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$, and the class of neural network classifiers of a given architecture, given by $\mathcal{H}_{\text{NET}} = \{x \mapsto f(x, w) : w \in W\}$, where $f$ represents a neural network architecture and $W$ encompasses the set of all conceivable choices of weights for this specific architecture. With these examples in mind, the problem of data fitting can be formulated as the optimization problem (2.1).

$$\min_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{h(x_i)\neq y_i}. \tag{2.1}$$

This methodology gives rise to perhaps the simplest learning algorithm: Empirical Risk Minimization over $\mathcal{H}$ (ERM-$\mathcal{H}$). Given the training sample $(x_1, y_1), ..., (x_m, y_m)$, it consists in finding

$$\hat{h}_m \in \text{argmin}_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{h(x_i)\neq y_i}. \tag{2.2}$$

This setup motivates a lot of interesting questions within multiple scientific fields, including optimization [SNW11], theoretical computer science [Val84], and mathematical statistics. In this thesis, we delve into the realm of statistical learning theory [Vap10] to understand when ERM-$\mathcal{H}$ can be expected to successfully learn from the available data. In particular, we will see that the success of learning hinges on three things:

1. On the ability of the chosen model class to approximate the desired pattern, i.e., on there being $h \in \mathcal{H}$ that often produces the right classifications;

2. On the complexity of the chosen model class, as measured by *VC dimension* (Definition 2.4.4) or *Rademacher complexity* (Definition 2.4.1);

3. On the number of samples available for training.

In machine learning applications, practitioners often have control over both the complexity of the model class and the number of samples. As such, this chapter provides a sound framework for thinking about phenomena like generalization, overfitting, and underfitting. In particular, given a choice of $\delta \in (0, 1)$, our analysis yields upper bounds on the generalization error of ERM-$\mathcal{H}$ of the following form:

$$\mathbb{P}\left(\widehat{\text{gen}}_m(\hat{h}_m) \leq \varepsilon(\text{Complexity}(\mathcal{H}), \delta, m)\right) \geq 1 - \delta, \tag{2.3}$$

where $\hat{h}_m$ is the same as in (2.2), $\widehat{\text{gen}}$ is formally introduced in Definition 2.1.2, and $\varepsilon \to 0$ as $m \to \infty$. These bounds enable us to characterize overfitting as attempting to learn a pattern using fewer samples than the required for the model class complexity.

discrepancy between the model's prediction $h(x)$ and the ground truth $y$. In this chapter, we stick with the natural choice of $L(h, x, y) = \mathbf{1}_{h(x)\neq y}$, which captures the notion of accuracy.

Additionally, from the first of the tree conditions above, underfitting can be seen as trying to learn with a model class that lacks the richness needed to capture the desired pattern.

As a note, the generalization bounds discussed in this chapter have significantly shaped the development of machine learning. In particular[3], margin-based bounds played a crucial role in influencing max-margin methodologies such as Support Vector Machines (SVMs) [ABR64; BGV92], which remained state-of-the-art until the early 2010s when deep neural networks emerged as dominant players [KSH12]. Unfortunately, the generalization ability of deep neural networks cannot be explained by the theory of this chapter, as these models don't respect the principle of capacity control, usually having orders of magnitude more parameters than data [Zha+16]. Understanding the mechanisms that affect the generalization of deep neural networks is still an active area of research [Aro+19; DR17; Ger+24].

This chapter is organized as follows. We begin in Section 2.1 by introducing the standard mathematical model for binary classification and discussing the concepts of learning and generalization within this framework. In Section 2.2, we deepen our discussion on empirical risk minimization, laying ground with a key lemma (Lemma 2.2.1) that will be pivotal throughout the text. In Section 2.3, we introduce tools from probability theory, in particular the concentration inequalities of *Chernoff-Hoeffding* (Theorem 2.3.1) and *McDiarmid* (Theorem 2.3.3), which we then apply to obtain non-asymptotic upper bounds on the *estimation error* (defined in Section 2.1) of ERM-$\mathcal{H}$ when using a finite class $\mathcal{H}$. Under the *inductive bias* assumption that will be introduced in Section 2.1, this entails that learning with ERM-$\mathcal{H}$, $\mathcal{H}$ finite, is always possible given enough data. In Section 2.4, we generalize these results to infinite classes $\mathcal{H}$. We begin by employing concentration inequality machinery to upper bound the estimation error in terms of *Rademacher complexity* (Definition 2.4.1). Although the bounds are tight, they are not easily computable in practice. As such, we upper bound the Rademacher complexity with yet another complexity measure: the *VC dimension* (Definition 2.4.4). Building on this, we establish the fundamental result of the chapter, which states that learning is only feasible when using classes that have finite VC dimension (see Theorem 2.4.4).

To navigate this chapter effectively, readers should be familiar with measure-theoretic probability, as covered in [AL06]. Additionally, the text presupposes an informal understanding of basic machine learning concepts such as binary classification and risk (or loss) minimization, as presented in [DFO20]. For a more comprehensive understanding, readers can refer to [MRT18, Chapters 1-3].

[3] For a more recent example in the context of learning dynamical systems, see [Kos+22].

## 2.1 Binary classification

The purpose of this section is to present a mathematical model for binary classification and to formalize the notions of learning and generalization. In essence, we aim to develop a theoretical framework for asserting the future performance of classifiers, which clearly involves introducing modeling assumptions on both the training and test data points. This sort of data modeling problem is typical in statistics, and the standard solution is to assume the existence of a data-generating process from which all points under consideration are drawn.

Given feature and label spaces $\mathcal{X}$ and $\mathcal{Y} = \{-1, 1\}$, the simplest and most frequently adopted assumption is that points are drawn independently and identically distributed (iid) from an unknown but fixed probability distribution $\mathbb{P}$ over $\mathcal{X}$. To accommodate more realistic scenarios, where $y$ is not solely determined by $x$, it is common to consider a probability distribution over $\mathcal{X} \times \mathcal{Y}$. This is the framework we shall adhere to in this thesis. As a note, the present framework can be extended to encompass interdependence between samples and a non-stationary data-generating process. For further details, refer to [Dar+15].

Having motivated our mathematical framework, we now describe it in more detail. Let $\mathcal{X}$ be a feature space and $\mathcal{Y} = \{-1, 1\}$. Consider a sigma-algebra $\mathcal{A}$ on $\mathcal{X}$ so that $\mathcal{X} \times \mathcal{Y}$ is equipped with the product sigma-algebra $\mathcal{A} \times \mathcal{P}(\mathcal{Y})$ and $\mathcal{P}(\cdot)$ denotes the powerset operation. Henceforth, we shall consider a fixed but unknown probability measure $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$. In this context, learning consists in finding a classifier $h$ with small risk (Definition 2.1.1) and a model that generalizes is one with small generalization error as in Definition 2.1.2.

**Definition 2.1.1** (**Risk**). *Let $h$ be a classifier, then we define its risk to be*

$$R(h) := \mathbb{P}(h(x) \neq y), \tag{2.4}$$

*Furthermore, given a sample $(x_1, y_1), ..., (x_m, y_m) \overset{iid}{\sim} \mathbb{P}$, the empirical risk of the classifier $h$ with respect to the sample is given by*

$$\hat{R}_m(h) := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq y_i}. \tag{2.5}$$

**Remark 2.1.1.** *Note that*

$$\begin{aligned}
\mathbb{E}_{(x_1,y_1),...,(x_m,y_m) \overset{iid}{\sim} \mathbb{P}} \left[ \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq y_i} \right] &= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{(x_1,y_1),...,(x_m,y_m) \overset{iid}{\sim} \mathbb{P}} \left[ \sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq y_i} \right] && (2.6)\\
&= \mathbb{E}\Big[ \mathbf{1}_{h(x_1) \neq y_1} \Big] && (2.7)\\
&= R(h). && (2.8)
\end{aligned}$$

**Remark 2.1.2.** *For expressions such as* (2.4) *to make sense, we have to assume that model classes* $\mathcal{H}$ *are always comprised of measurable functions with respect to the sigma-algebras* $\mathcal{A}$ *and* $\mathcal{P}(\{-1,1\})$*.*

**Definition 2.1.2 (Generalization error).** *Let* $h$ *be a classifier and* $(x_1, y_1), ..., (x_m, y_m) \overset{iid}{\sim} \mathbb{P}$ *be a sample. Then, the generalization error of* $h$ *with respect to the sample is given by*

$$\hat{\text{gen}}_m(h) = R(h) - \hat{R}_m(h). \tag{2.9}$$

**Remark 2.1.3.** *In section 2.2 we will see that the tasks of learning and minimizing the generalization error can be approached simultaneously (see Remark* 2.2.2*).*

**Remark 2.1.4.** *As we are not assuming the existence of a functional relation between* $x$ *and* $y$*, there may not necessarily be a classifier that achieves zero risk for any given* $\mathbb{P}$*. Consider, for instance, the extreme case where* $y$ *is entirely independent of* $x$*; e.g.,* $y$ *is determined by tossing an unbiased coin. In this scenario, the risk of any classifier is* $1/2$*.*

**Remark 2.1.5.** *Notice it is not possible to find a classifier that achieves a smaller risk than* $h(x) = \text{sign}(\eta(x) - 1/2)$*, where* $\eta(x) = \mathbb{P}(y = +1 \mid X = x)$*. Indeed, we may write*

$$\begin{aligned}
\mathbb{P}(h(x) \neq y) &= \mathbb{E}_{(x,y)\sim\mathbb{P}}[\mathbf{1}_{h(x)\neq y}] && (2.10)\\
&= \mathbb{E}_x\mathbb{E}_y[\mathbf{1}_{h(x)\neq 1}\mathbf{1}_{y=+1} + \mathbf{1}_{h(x)\neq -1}\mathbf{1}_{y=-1}] && (2.11)\\
&= \mathbb{E}_x[\eta(x)\mathbf{1}_{h(x)=-1} + (1-\eta(x))\mathbf{1}_{h(x)=1}]. && (2.12)
\end{aligned}$$

*The expression* (2.12) *is minimized by choosing a classifier that is* 1 *when* $\eta(x) \geq 1/2$ *and* $-1$ *otherwise. This classifier is known as the* Bayes classifier *and its risk is denoted by* $R^* = R(h^*)$*.*

In light of Remark 2.1.4, it is evident that the extent to which learning can occur is ultimately dependent on the underlying distribution $\mathbb{P}$. For easier problems, a simpler class $\mathcal{H}$ may be expressive enough for

$$\inf_{h\in\mathcal{H}} R(h), \tag{2.13}$$

to be small, whereas for harder problems, a more complex class is required to maintain the same level of performance. In order to keep the generality of the theory, it is essential that we operate under the assumption that the given problem is well modelled by the chosen class $\mathcal{H}$. This assumption is known as a *flat inductive bias* towards the class $\mathcal{H}$. For further details on its necessity, readers are encouraged to consult [Ger+24, Chapter 1] and references therein. Under this assumption, learning is reduced to minimizing the *estimation error*

$$R(h) - \inf_{h\in\mathcal{H}} R(h). \tag{2.14}$$

From Section 2.2 onwards, we focus exclusively on studying the *estimation error* of ERM-$\mathcal{H}$.

## 2.2 Empirical risk minimization

In the previous section we introduced the standard model for binary classification. In this section, we begin the analysis of the empirical risk minimization algorithm (ERM-$\mathcal{H}$) (Definition 2.2.1). In particular, we present a key result that allows us to bound the *estimation error* of ERM-$\mathcal{H}$ by the largest generalization error realized by a classifier inside $\mathcal{H}$. That is, by

$$\hat{\text{err}}_m(\mathcal{H}) = \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_m(h)|. \tag{2.15}$$

The first application of this bound is outlined in Proposition 2.2.2, which shows that the *estimation error* of ERM-$\mathcal{H}$ asymptotically goes to zero as $m \to \infty$ when $\mathcal{H}$ is finite. We will make substantial improvements to this bound after introducing the necessary concentration inequalities in Section 2.3.

**Definition 2.2.1 (ERM-$\mathcal{H}$).** *Given a model class $\mathcal{H}$ and a sample $(x_1, y_1), ..., (x_m, y_m) \overset{iid}{\sim} \mathbb{P}$, the* ERM-$\mathcal{H}$ *algorithm returns $\hat{h}_m$ such that*

$$\hat{h}_m \in \text{argmin}_{h \in \mathcal{H}} \hat{R}_m(h), \tag{2.16}$$

*where $\hat{R}_m(h) = \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{h(x_i) \neq y_i}$ is the empirical risk.*

**Remark 2.2.1.** *Achieving the minimum in* (2.16) *may not be possible. However, for our purposes, it suffices to approximately solve the optimization problem, as seen in Lemma* 2.2.1.

**Lemma 2.2.1.** *Let $\mathcal{H}$ be any collection of models and define*

$$\hat{\text{err}}_m(\mathcal{H}) = \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_m(h)|, \tag{2.17}$$

*where $(x_1, y_1), ..., (x_m, y_m) \overset{iid}{\sim} \mathbb{P}$ , then for $\hat{h}_m$ from* (2.16)*, we have*

$$R(\hat{h}_m) - \inf_{h \in \mathcal{H}} R(h) \leq 2\hat{\text{err}}_m(\mathcal{H}). \tag{2.18}$$

*Additionally, if $\hat{h}_m$ only approximately solves* (2.16)*; i.e., if $\hat{R}_m(\hat{h}_m) \leq \hat{R}_m(h) + \varepsilon$ for every $h \in \mathcal{H}$, then we have*

$$R(\hat{h}_m) - \inf_{h \in \mathcal{H}} R(h) \leq 2\hat{\text{err}}_m(\mathcal{H}) + \varepsilon. \tag{2.19}$$

*Proof.* In the case $\hat{h}_m$ minimizes the empirical risk, we have that

$$\hat{R}_m(\hat{h}_m) \leq \hat{R}_m(h), \quad \forall h \in \mathcal{H}. \tag{2.20}$$

In particular, for $h_\epsilon \in \mathcal{H}$ such that $R(h_\epsilon) \leq \inf_{h\in\mathcal{H}} R(h) + \epsilon$, we also have $\hat{R}_m(\hat{h}_m) \leq \hat{R}_m(h_\epsilon)$. Moreover,

$$\begin{aligned} R(\hat{h}_m) &= \hat{R}_m(\hat{h}_m) + (R(\hat{h}_m) - \hat{R}_m(\hat{h}_m)) && (2.21)\\ &\leq \hat{R}_m(h_\epsilon) + \hat{\text{err}}_m(\mathcal{H}) && (2.22)\\ &= R(h_\epsilon) + (\hat{R}_m(h_\epsilon) - R(h_\epsilon)) + \hat{\text{err}}_m(\mathcal{H}) && (2.23)\\ &\leq \inf_{h\in\mathcal{H}} R(h) + \epsilon + 2\hat{\text{err}}_m(\mathcal{H}). && (2.24) \end{aligned}$$

As this is valid for all $\epsilon > 0$, it is also valid without it. Thus,

$$R(\hat{h}_m) - \inf_{h\in\mathcal{H}} R(h) \leq 2\hat{\text{err}}_m(\mathcal{H}), \tag{2.25}$$

as desired. If $\hat{h}_m$ only satisfies

$$\hat{R}_m(\hat{h}_m) \leq \hat{R}_m(h) + \varepsilon, \quad \forall h \in \mathcal{H}, \tag{2.26}$$

then (2.22) gets an extra $\varepsilon$, which propagates until the end to give us

$$R(\hat{h}_m) - \inf_{h\in\mathcal{H}} R(h) \leq 2\hat{\text{err}}_m(\mathcal{H}) + \varepsilon. \tag{2.27}$$

□

**Remark 2.2.2.** *The quantity* $\hat{\text{err}}_m(\mathcal{H})$ *appearing in Lemma* 2.2.1 *is crucial to statistical learning theory, as it provides a unified way to study the generalization error of any learning algorithm. Indeed, given a learning algorithm A that produces a classifier* $A(S) \in \mathcal{H}$ *given a sample* $S = (x_1, y_1), ..., (x_m, y_m)$*, it is evident that*

$$\hat{\text{gen}}_m(A(S)) \leq \hat{\text{err}}_m(\mathcal{H}). \tag{2.28}$$

*As such, any high-probability upper bound on* $\hat{\text{err}}_m(\mathcal{H})$ *immediately yields a high-probability upper bound on the generalization error of A. In Section 2.4 we will relate the complexity of* $\mathcal{H}$ *to* $\mathbb{E}[\hat{\text{err}}_m(\mathcal{H})]$ *and the rate by which* $\hat{\text{err}}_m(\mathcal{H})$ *converges to* $\mathbb{E}[\hat{\text{err}}_m(\mathcal{H})]$*, in the attempt to show that* $\hat{\text{err}}_m(\mathcal{H})$ *quickly converges to zero. Thus, we will show that both the generalization and estimation errors of ERM-*$\mathcal{H}$ *converge to zero.*

## Preliminary results on empirical risk minimization

In this short subsection, we give some early results on ERM-$\mathcal{H}$. Specifically, we consider a sequence of data points $(x_1, y_1), (x_2, y_2), ... \overset{iid}{\sim} \mathbb{P}$, and show that $\hat{R}_m(h)$ converges in probability to $R(h)$ as $m \to \infty$ (Proposition 2.2.1). Furthermore, we extend this reasoning to show that the *estimation error* of ERM-$\mathcal{H}$ converges to zero in probability for finite classes $\mathcal{H}$ (Proposition 2.2.2). For these results, we need the following version of the Weak Law of Large Numbers.

**Theorem 2.2.1** (**Weak Law of Large Numbers** [AL06, p. 238])**.** *Let $X_1, X_2, \ldots$ be an infinite sequence of i.i.d. random variables with mean $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \ldots = \mu$. Then, the empirical averages*

$$\hat{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i \tag{2.29}$$

*converge to the expected value $\mu$ in probability:*

$$\hat{X}_m \xrightarrow{\mathbb{P}} \mu, \quad \text{as } m \to \infty. \tag{2.30}$$

**Proposition 2.2.1.** *Given a classifier $h \in \mathcal{H}$, then*

$$\hat{R}_m(h) \xrightarrow{\mathbb{P}} R(h), \quad \text{as } m \to \infty. \tag{2.31}$$

*Proof.* Consider an infinite sample $\left\{(x_i, y_i)\right\}_{i=1}^{\infty} \overset{iid}{\sim} \mathbb{P}$, then the sequence $\left\{\mathbf{1}_{h(x_i)\neq y_i}\right\}_{i=1}^{\infty}$ is also i.i.d. as the map $L(x,y) = \mathbf{1}_{h(x)\neq y}$ is measurable from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. Thus, in view of Definition 2.1.1, by applying the **Weak Law of Large Numbers** (Theorem 2.2.1), we have that $\hat{R}_m(h) \xrightarrow{\mathbb{P}} R(h)$, as desired. □

**Proposition 2.2.2.** *Let $\mathcal{H}$ be finite, then*

$$R(\hat{h}_m) \xrightarrow{\mathbb{P}} \inf_{h\in\mathcal{H}} R(h). \tag{2.32}$$

*Proof.* By Lemma 2.2.1, we have that

$$\mathbb{P}\Big\{R(\hat{h}_m) - \inf_{h\in\mathcal{H}} R(h) > \epsilon\Big\} \leq \mathbb{P}\Big\{\widehat{\mathrm{err}}_m(\mathcal{H}) > \epsilon/2\Big\} \tag{2.33}$$

$$= \mathbb{P}\Big\{\max_{h\in\mathcal{H}} |R(h) - \hat{R}_m(h)| > \epsilon/2\Big\} \tag{2.34}$$

$$\leq \sum_{h\in\mathcal{H}} \mathbb{P}\Big\{|R(h) - \hat{R}_m(h)| > \epsilon/2\Big\}. \tag{2.35}$$

By taking the limit $m \to \infty$ and using Proposition 2.2.1 we get the desired result. □

## 2.3 Concentration inequalities

So far, we have given asymptotic results on the performance of ERM-$\mathcal{H}$. In practice, there are only finitely many data points and so non-asymptotic results become crucial. To address this, we rely on a series of results called *concentration inequalities.*

Our approach is self-contained and focused on learning theory. For a more comprehensive treatment, readers can refer to [BLM16].

We begin by introducing *Chernoff's bounding technique* [Che52]: suppose you are interested in bounding the tail event

$$\mathbb{P}\Big\{X \geq a\Big\}, \tag{2.36}$$

where $X$ is any random variable that has a moment-generating function (MGF). Let $M(t) = \mathbb{E}[e^{tX}]$ be its MGF, then the technique consists in applying *Markov's inequality* to get the MGF and then taking the infimum. Indeed, for any $t > 0$, we have

$$\mathbb{P}\Big\{X \geq a\Big\} = \mathbb{P}\Big\{e^{tX} \geq e^{ta}\Big\} \tag{2.37}$$
$$\leq M(t)e^{-ta}, \tag{2.38}$$

which gives

$$\mathbb{P}\Big\{X \geq a\Big\} \leq \inf_{t>0} M(t)e^{-ta}. \tag{2.39}$$

**Remark 2.3.1.** *One might wonder whether these bounds are any good, since they look so arbitrary. This is the content of Cramér's theorem on large deviations [CT18, Thm. 6], which shows that such bounds are optimal in the case where $X$ is a sum of i.i.d. random variables.*

Before applying this technique to our problems, we need the following technical lemma.

**Lemma 2.3.1 (Hoeffding** [Hoe63]**).** *Let $X : \Omega \to [a, b]$ be a random variable with MGF $M : (0, \infty) \to \mathbb{R}_+$ and satisfying $\mathbb{E}[X] = 0$, then*

$$\log(M(t)) \leq \frac{t^2(b-a)^2}{8}. \tag{2.40}$$

With this, we can prove the following.

**Theorem 2.3.1 (Chernoff-Hoeffding** [Hoe63, Thm. 2]**).** *For $i = 1, ..., m$, let $T_i : \Omega \to [a_i, b_i]$ be independent random variables defined on a common probability space $\Omega$. Then, for every $\epsilon > 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^{m}(T_i - \mathbb{E}[T_i]) > \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right) \tag{2.41}$$
$$\mathbb{P}\left\{\left|\sum_{i=1}^{m}(T_i - \mathbb{E}[T_i])\right| > \epsilon\right\} \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right). \tag{2.42}$$

*Proof.* Without loss of generality, suppose that $\mathbb{E}[T_i] = 0$ (otherwise, consider $\tilde{T}_i = T_i - \mathbb{E}[T_i]$). Now, we apply *Chernoff's technique*:

$$\mathbb{P}\left\{\sum_{i=1}^{m} T_i > \epsilon\right\} = \mathbb{P}\left\{e^{t\sum_{i=1}^{m} T_i} > e^{t\epsilon}\right\} \tag{2.43}$$

$$\leq \mathbb{E}\left[e^{t\sum_{i=1}^{m} T_i}\right]e^{-t\epsilon} \tag{2.44}$$

$$= \mathbb{E}\left[\prod_{i=1}^{m} e^{tT_i}\right]e^{-t\epsilon} \tag{2.45}$$

$$= \prod_{i=1}^{m} \mathbb{E}\left[e^{tT_i}\right]e^{-t\epsilon} \tag{2.46}$$

$$\leq \exp\left(\frac{t^2}{8}\sum_{i=1}^{m}(b_i - a_i)^2 - t\epsilon\right), \tag{2.47}$$

where the last inequality follows from Lemma 2.3.1. Minimizing on $t$, we get that $t = \frac{4\epsilon}{\sum_{i=1}^{m}(b_i-a_i)^2}$. Hence,

$$\mathbb{P}\left\{\sum_{i=1}^{m} T_i > \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right), \tag{2.48}$$

as desired. The second inequality follows from the first by considering the random variables $-T_i : \Omega \to [-b_i, -a_i]$. The first inequality then implies:

$$\mathbb{P}\left\{\sum_{i=1}^{m}(-T_i + \mathbb{E}[T_i]) > \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right). \tag{2.49}$$

The *union bound* finishes the proof:

$$\mathbb{P}\left\{\left|\sum_{i=1}^{m}(T_i - \mathbb{E}[T_i])\right| > \epsilon\right\} = \mathbb{P}\left\{\sum_{i=1}^{m}(T_i - \mathbb{E}[T_i]) > \epsilon \bigvee \sum_{i=1}^{m}(-T_i + \mathbb{E}[T_i]) > \epsilon\right\} \tag{2.50}$$

$$\leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right) \tag{2.51}$$

□

Chernoff-Hoeffding's inequality (Theorem 2.3.1) enables us to characterize the rate of convergence of $R(\hat{h}_m)$ to $\inf_{h\in\mathcal{H}} R(h)$, as the following two results show.

**Corollary 2.3.1.** *Let $\mathcal{G}$ be a finite family of measurable functions $g : \mathcal{Z} \to [0, 1]$ and $z_1, ..., z_m \overset{iid}{\sim} \mathbb{P}$. Then, for every $\epsilon > 0$, we have*

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathbb{E}_{z\sim\mathbb{P}}[g(z)]\right| > \epsilon\right\} \leq 2\#\mathcal{G}\exp\Big(-2m\epsilon^2\Big). \tag{2.52}$$

*Proof.* Let $T_i = g(z_i)$. Note that the $g(z_i)$ are independent and that $b_i - a_i = 1$ since the image of $g$ is contained in $[0, 1]$. Thus, by **Chernoff-Hoeffding's inequality** (Theorem 2.3.1), we have

$$\mathbb{P}\bigg\{\sup_{g\in\mathcal{G}}\bigg|\frac{1}{m}\sum_{i=1}^{m}g(z_i)-\mathbb{E}_{z\sim\mathbb{P}}[g(z)]\bigg|>\epsilon\bigg\}=\mathbb{P}\bigg\{\exists g\in\mathcal{G}:\bigg|\frac{1}{m}\sum_{i=1}^{m}g(z_i)-\mathbb{E}_{z\sim\mathbb{P}}[g(z)]\bigg|>\epsilon\bigg\} \tag{2.53}$$

$$\leq\sum_{g\in\mathcal{G}}2\exp\Big(-2m\epsilon^2\Big) \tag{2.54}$$

$$=2\#\mathcal{G}\exp\Big(-2m\epsilon^2\Big), \tag{2.55}$$

as desired. □

**Theorem 2.3.2.** *Let $\mathcal{H}$ be finite, then*

$$\mathbb{P}\bigg\{R(\hat{h}_m)-\inf_{h\in\mathcal{H}}R(h)>\epsilon\bigg\}\leq 2\#\mathcal{H}\exp\left(\frac{-m\epsilon^2}{2}\right), \tag{2.56}$$

*which can be expressed in the form of a high-probability bound: for every $\delta\in(0,1)$, with probability at least $1-\delta$:*

$$R(\hat{h}_m)-\inf_{h\in\mathcal{H}}R(h)\leq\sqrt{\frac{2\log\left(\frac{2\#\mathcal{H}}{\delta}\right)}{m}}. \tag{2.57}$$

*Proof.* Let $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ and $\mathcal{G}=\{(x,y)\mapsto\mathbf{1}_{h(x)\neq y}:h\in\mathcal{H}\}$, then, by Lemma 2.2.1 and Corollary 2.3.1,

$$\mathbb{P}\bigg\{R(\hat{h}_m)-\inf_{h\in\mathcal{H}}R(h)>\epsilon\bigg\}\leq\mathbb{P}\bigg\{\max_{h\in\mathcal{H}}|R(h)-\hat{R}_m(h)|>\epsilon/2\bigg\} \tag{2.58}$$

$$=\mathbb{P}\bigg\{\max_{g\in\mathcal{G}}\bigg|\mathbb{E}[g(z))]-\frac{1}{m}\sum_{i=1}^{m}g(z_i)\bigg|>\epsilon/2\bigg\} \tag{2.59}$$

$$\leq 2\#\mathcal{H}\exp\left(\frac{-m\epsilon^2}{2}\right). \tag{2.60}$$

To get the generalization bound, set the rhs to $\delta$ and solve for $\epsilon$:

$$\delta=2\#\mathcal{H}\exp\left(\frac{-m\epsilon^2}{2}\right) \tag{2.61}$$

$$\frac{\delta}{2\#\mathcal{H}}=\exp\left(\frac{-m\epsilon^2}{2}\right) \tag{2.62}$$

$$\log\left(\frac{\delta}{2\#\mathcal{H}}\right)=\frac{-m\epsilon^2}{2} \tag{2.63}$$

$$\sqrt{\frac{2\log\left(\frac{2\#\mathcal{H}}{\delta}\right)}{m}}=\epsilon. \tag{2.64}$$

□

For the next section, we will need a stronger version of Chernoff-Hoeffding's inequality (Theorem 2.3.1) known as McDiarmid's inequality (Theorem 2.3.3).

**Theorem 2.3.3** (**McDiarmid** [Doo40])**.** *For $i = 1, ..., m$, let $X_i : \Omega \to \mathcal{X}, X_i \sim \mathbb{P}_i$, be independent random variables defined on a common probability space $\Omega$. Additionally, assume there are $c_1, c_2, ..., c_m > 0$ such that $f : \mathcal{X}^m \to \mathbb{R}$ satisfies the bounded differences condition:*

$$\Big| f(x_1, ..., x_i, ..., x_m) - f(x_1, ..., x_i', ..., x_m) \Big| \leq c_i, \tag{2.65}$$

*for all $i \in [m]$ and $x_1, ..., x_m, x_i' \in \mathcal{X}$. Denote by $X_1^m$ the vector $(X_1, ..., X_m)$. Then, for all $\varepsilon > 0$,*

$$\mathbb{P}\Big\{ f(X_1^m) - \mathbb{E}[f(X_1^m)] > \varepsilon \Big\} \leq \exp\left( \frac{-2\varepsilon^2}{\sum_{i=1}^m c_i^2} \right). \tag{2.66}$$

*Proof.* We begin with *Chernoff's bounding technique*: for $t > 0$, we have

$$\begin{aligned} \mathbb{P}\Big\{ f(X_1^m) - \mathbb{E}[f(X_1^m)] > \varepsilon \Big\} &= \mathbb{P}\Big\{ e^{t(f(X_1^m) - \mathbb{E}[f(X_1^m)])} > e^{t\varepsilon} \Big\} && (2.67) \\ &\leq \inf_{t>0} e^{-t\varepsilon} \mathbb{E}\Big[ e^{t(f(X_1^m) - \mathbb{E}[f(X_1^m)])} \Big] && (2.68) \end{aligned}$$

We establish control over the MGF in (2.68) using induction. The base case ($m = 1$) follows from **Hoeffding's lemma** (Lemma 2.3.1). Indeed, define

$$A = \inf_{x \in \mathcal{X}} f(x), \quad B = \sup_{x \in \mathcal{X}} f(x). \tag{2.69}$$

The random variable $f(X_1^1)$ takes values on $[A, B]$, where $B - A \leq c_1$, so that, by Lemma 2.3.1, we have

$$\mathbb{E}\Big[ e^{t(f(X_1^1) - \mathbb{E}[f(X_1^1)])} \Big] \leq \exp\left( \frac{t^2 c_1^2}{8} \right). \tag{2.70}$$

For $m > 1$, consider the following definitions

$$\begin{aligned} f_{m-1}(x_1^{m-1}) &= \int_{\mathcal{X}} f(x_1, ..., x_{m-1}, x_m) d\mathbb{P}_m(x_m), && (2.71) \\ \Delta_m(x_1^m) &= f(x_1^m) - f_{m-1}(x_1^{m-1}). && (2.72) \end{aligned}$$

Notice that $\mathbb{E}[f(X_1^m)] = \mathbb{E}[f_{m-1}(X_1^{m-1})]$ so that we can write

$$\mathbb{E}\Big[ e^{t(f(X_1^m) - \mathbb{E}[f(X_1^m)])} \Big] = \mathbb{E}\Big[ e^{t\Delta_m(X_1^m)} e^{t(f_{m-1}(X_1^{m-1}) - \mathbb{E}[f_{m-1}(X_1^{m-1})])} \Big]. \tag{2.73}$$

By fixing the values of $x_1, ..., x_{m-1}$ and integrating first with respect to $x_m$, we get

$$\int_{\mathcal{X}} e^{t\Delta_m(x_1^m)} e^{t(f_{m-1}(x_1^{m-1})-\mathbb{E}[f_{m-1}(x_1^{m-1})])} d\mathbb{P}_m(x_m)$$
$$= e^{t(f_{m-1}(x_1^{m-1})-\mathbb{E}[f_{m-1}(x_1^{m-1})])} \mathbb{E}_{x_m\sim\mathbb{P}_m}\left[e^{t\Delta_m(x_1^m)}\right]. \quad (2.74)$$

If we let $g : \mathcal{X} \to \mathbb{R}$ be given by $g(x) = f(x_1, ..., x_{m-1}, x)$, then its clear that $g$ satisfies the bounded differences condition (2.65) with constant $c_m$. Moreover, by a reduction to the base case ($m = 1$), it is possible to see that

$$\mathbb{E}_{x_m\sim\mathbb{P}_m}\left[e^{t\Delta_m(x_1^m)}\right] = \mathbb{E}\left[e^{t(g(X_m^m)-\mathbb{E}(g(X_m^m))}\right] \leq \exp\left(\frac{t^2c_m^2}{8}\right). \quad (2.75)$$

Now, if we integrate with respect to the remaining variables $x_1, ..., x_{m-1}$, we obtain

$$\mathbb{E}\left[e^{t(f(X_1^m)-\mathbb{E}[f(X_1^m)])}\right] \leq \mathbb{E}\left[e^{t(f_{m-1}(X_1^{m-1})-\mathbb{E}[f_{m-1}(X_1^{m-1})])}\right] \exp\left(\frac{t^2c_m^2}{8}\right). \quad (2.76)$$

By noting that $f_{m-1}$ respects the bounded differences condition (2.65), we apply induction to get

$$\mathbb{E}\left[e^{t(f(X_1^m)-\mathbb{E}[f(X_1^m)])}\right] \leq \exp\left(\sum_{i=1}^{m} \frac{t^2c_i^2}{8}\right). \quad (2.77)$$

Analogously to the proof of **Chernoff-Hoeffding's inequality** (Theorem 2.3.1), we substitute the last expression into (2.68) and minimize with respect to $t > 0$. This finishes the proof. □

## 2.4 Empirical risk minimization over infinite classes

Although Theorem 2.3.2 is already a great improvement over Proposition 2.2.2, it continues to be limited by the fact that $\mathcal{H}$ must be finite. To overcome this limitation, in this section we delve deeper into the analysis of the *empirical process*[4] from Lemma 2.2.1:

$$\hat{\text{err}}_m(\mathcal{H}) = \sup_{h\in\mathcal{H}} \left|R(h) - \hat{R}_m(h)\right|, \quad (2.78)$$

which we will henceforth write as

$$\sup_{g\in\mathcal{G}} \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathbb{E}[g(z)]\right|, \quad (2.79)$$

[4] In probability theory, empirical processes are families of random variables of the form $\{\sum_{i=1}^m f(x_i) - \int f d\mathbb{P} : f \in \mathcal{F}\}$, where $\mathcal{F}$ is a collection of measurable functions $f : \mathcal{X} \to \mathbb{R}$. In the sub-field of empirical process theory, it is common to study the convergence properties of such random variables and to characterize for which classes $\mathcal{F}$ different modes of convergence can be shown. In this context, the analysis of this section (specially Theorem 2.4.4) can be seen as an analysis of the uniform convergence of the empirical process $||\mathbb{P}_n - \mathbb{P}||_{\mathcal{G}} = \sup_{g\in\mathcal{G}} |\frac{1}{m}\sum_{i=1}^m g(z_i) - \mathbb{E}[g(z)]|$. For more information on this fascinating connection, consult [VW23].

where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{G} = \{(x, y) \mapsto \mathbf{1}_{h(x) \neq y} : h \in \mathcal{H}\}$, and $z_1, ..., z_m \overset{iid}{\sim} \mathbb{P}$. In particular, we note that the function

$$f(z_1, ..., z_m) = \sup_{g \in \mathcal{G}} \left| \frac{1}{m} \sum_{i=1}^{m} g(z_i) - \mathbb{E}[g(z)] \right|. \tag{2.80}$$

satisfies the bounded differences condition (2.65), which implies that the random variable $\hat{\text{err}}_m(\mathcal{H})$ quickly concentrates around its mean $\text{err}_m(\mathcal{H}) = \mathbb{E}[\hat{\text{err}}_m(\mathcal{H})]$ (see Corollary 2.4.1).

**Corollary 2.4.1.** *Given a class $\mathcal{H}$ and a sample $z_1, ..., z_m \overset{iid}{\sim} \mathbb{P}$, we have*

$$\mathbb{P}\left\{ \hat{\text{err}}_m(\mathcal{H}) > \epsilon + \mathbb{E}[\hat{\text{err}}_m(\mathcal{H})] \right\} \leq \exp\left( -\frac{m\epsilon^2}{2} \right). \tag{2.81}$$

As such, we shift our analysis towards understanding the simpler quantity $\text{err}_m(\mathcal{H})$ in terms of the capacity of the class $\mathcal{H}$ of generating distinct classifications. Specifically, we start in Subsection 2.4 by developing the notion of *Rademacher complexity* (Definition 2.4.1) and showing that it can be used to bound $\text{err}_m(\mathcal{H})$ via a *Symmetrization* result (Lemma 2.4.1). Although the obtained bound is sharp, it is not easily computable in practice, as the Rademacher complexity depends on the unkown distribution $\mathbb{P}$. This motivates the introduction of a simpler, distribution-free, notion of complexity: the VC dimension. In Subsection 2.4, we develop the concept and show that it can remarkably be used to characterize the family of classes $\mathcal{H}$ for which $\text{err}_m(\mathcal{H}) \to 0$ as $m \to \infty$, uniformly over all distributions $\mathbb{P}$ (see Theorem 2.4.4).

*Proof of Corollary.* Fix $z_1, ..., z_i, z_i', ..., z_m$ and define $f(z_1, ..., z_m) = \hat{\text{err}}_m(\mathcal{H})$. For conciseness, let $C(g) = \frac{1}{m} \sum_{j \neq i}^{m} g(z_j) - \mathbb{E}[g(z)]$ so that we can write

$$\begin{aligned} &|f(z_1, ..., z_i, ..., z_m) - f(z_1, ..., z_i', ..., z_m)| \\ &\qquad = \left| \sup_{g \in \mathcal{G}} \left| \frac{1}{m} g(z_i) + C(g) \right| - \sup_{g \in \mathcal{G}} \left| \frac{1}{m} g(z_i') + C(g) \right| \right|, \end{aligned} \tag{2.82}$$

We proceed to show that (2.82) $\leq \frac{2}{m}$. Indeed, suppose without loss of generality that $\sup_{g \in \mathcal{G}} \left| \frac{1}{m} g(z_i) + C(g) \right| \geq \sup_{g \in \mathcal{G}} \left| \frac{1}{m} g(z_i') + C(g) \right|$. Then,

$$\left|\sup_{g\in\mathcal{G}}\left|\frac{1}{m}g(z_i)+C(g)\right|-\sup_{g\in\mathcal{G}}\left|\frac{1}{m}g(z_i')+C(g)\right|\right| \tag{2.83}$$

$$=\sup_{g\in\mathcal{G}}\left|\frac{1}{m}g(z_i)+C(g)\right|-\sup_{g\in\mathcal{G}}\left|\frac{1}{m}g(z_i')+C(g)\right| \tag{2.84}$$

$$\leq \frac{1}{m}+\sup_{g\in\mathcal{G}}|C(g)|-\sup_{g\in\mathcal{G}}\left|\frac{1}{m}g(z_i')+C(g)\right| \tag{2.85}$$

$$\leq \frac{1}{m}+\sup_{g\in\mathcal{G}}|C(g)|-\left(-\frac{1}{m}+\sup_{g\in\mathcal{G}}|C(g)|\right) \tag{2.86}$$

$$\leq \frac{2}{m}. \tag{2.87}$$

Now, let $X=(z_1,z_2,...,z_n)\overset{iid}{\sim}\mathbb{P}$. By **McDiarmid's inequality** (Theorem 2.3.3), we have that

$$\mathbb{P}\bigg\{\hat{\text{err}}_m(\mathcal{H})>\epsilon+\mathbb{E}[\hat{\text{err}}_m(\mathcal{H})]\bigg\}\leq\exp\left(-\frac{m\epsilon^2}{2}\right), \tag{2.88}$$

as desired. □

## Rademacher complexity and symmetrization

The Rademacher complexity serves as measure of richness for function classes $g:\mathcal{Z}\to[a,b]$ (see Definition 2.4.1). Although the notion applies to any such collection of functions, we will only consider those of the form $\mathcal{G}=\{(x,y)\mapsto\mathbf{1}_{h(x)\neq y}:h\in\mathcal{H}\}$, where $\mathcal{H}$ is a model class. In this context, it can be understood as the extent to which a hypothesis class $\mathcal{H}$ is able to fit random noise. Specifically, more complex classes should be able to produce more classification vectors $(h(x_1),...,h(x_m))$, and as such should more easily be able to maximize $\langle(\sigma_1,...,\sigma_m),(g(z_1),...,g(z_m))\rangle$, where $(\sigma_1,...,\sigma_m)\in\{-1,1\}^m$ is an arbitrary labeling of the data $(x_1,...,x_m)$. For a more comprehensive understanding, readers are encouraged to consult [MRT18, Chapter 3] and the original papers [BBL02; Kol01; KP00].

**Definition 2.4.1** (**Empirical Rademacher complexity**). *Let $\mathcal{G}$ be a family of functions from $\mathcal{Z}$ to $[a,b]$ and let $z_1,...,z_m\overset{iid}{\sim}\mathbb{P}$ be a sample of size $m$, then the empirical Rademacher complexity of $\mathcal{G}$ with respect to the sample is given by:*

$$\hat{\mathfrak{R}}_m(\mathcal{G})=\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^m\sigma_i g(z_i)\right|\right], \tag{2.89}$$

*where the expectation is taken only with respect to the* Rademacher random variables $\sigma_1,...,\sigma_m\overset{iid}{\sim}\mathbf{Unif}\{-1,+1\}$.

**Definition 2.4.2** (**Rademacher complexity**). *Let $\mathcal{G}$ be a family of functions from $\mathcal{Z}$ to $[a,b]$ and let $z_1,...,z_m\overset{iid}{\sim}\mathbb{P}$, then the Rademacher complexity of $\mathcal{G}$ is given by:*

$$\mathfrak{R}_m(\mathcal{G}) = \mathbb{E}\Big[\hat{\mathfrak{R}}_m(\mathcal{G})\Big]. \tag{2.90}$$

The following result relates our original problem with the notion of Rademacher complexity.

**Lemma 2.4.1** (**Symmetrization**). *Let* $z_1, ..., z_m \overset{iid}{\sim} \mathbb{P}$ *and* $\sigma_1, ..., \sigma_m \overset{iid}{\sim}$ **Unif**$\{-1, +1\}$ *be independent random variables, then*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \Big(g(z_i) - \mathbb{E}_{z\sim\mathbb{P}}[g(z)]\Big)\right|\right] \le 2\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \sigma_i g(z_i)\right|\right]. \tag{2.91}$$

*Proof.* We can assume that $(\Omega, \mathcal{F}, \mathbb{P})$ also supports $z'_1, ..., z'_m \overset{iid}{\sim} \mathbb{P}$ independent from the other random variables. Then, if we denote $(z_1, ..., z_m)$ by $z_1^m$,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \Big(g(z_i) - \mathbb{E}_{z\sim\mathbb{P}}[g(z)]\Big)\right|\right] = \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\mathbb{E}\left[\sum_{i=1}^m \Big(g(z_i) - g(z'_i)\Big) \;\Big|\; z_1^m\right]\right|\right] \tag{2.92}$$

$$\le \mathbb{E}\left[\sup_{g\in\mathcal{G}}\mathbb{E}\left[\left|\sum_{i=1}^m \Big(g(z_i) - g(z'_i)\Big)\right| \;\Big|\; z_1^m\right]\right] \tag{2.93}$$

$$\le \mathbb{E}\left[\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \Big(g(z_i) - g(z'_i)\Big)\right| \;\Big|\; z_1^m\right]\right] \tag{2.94}$$

$$= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \Big(g(z_i) - g(z'_i)\Big)\right|\right] \tag{2.95}$$

$$= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \sigma_i\Big(g(z_i) - g(z'_i)\Big)\right|\right] \tag{2.96}$$

$$\le \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \sigma_i g(z_i)\right| + \left|\sum_{i=1}^m \sigma_i g(z'_i)\right|\right] \tag{2.97}$$

$$\le \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \sigma_i g(z_i)\right| + \sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \sigma_i g(z'_i)\right|\right] \tag{2.98}$$

$$= 2\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^m \sigma_i g(z_i)\right|\right]. \tag{2.99}$$

The second line is an application of **Jensen's inequality** and the fifth involves the introduction of **Rademacher random variables** (see Definition 2.4.1). The expectation doesn't change because, given $\sigma_1^m$, the effect of each individual $\sigma_i$ just corresponds to swapping (or not) $z_i$ and $z'_i$. As $z_i$ and $z'_i$ come from the same distribution, this fixed swap doesn't change the distribution of the samples. □

We now put everything together to obtain the following fundamental result.

**Theorem 2.4.1.** *Let $\mathcal{H}$ be a hypothesis class. Then, with probability at least $1-\delta$, we have the following bounds on the* estimation error *of ERM-$\mathcal{H}$:*

$$R(\hat{h}_m) - \inf_{h\in\mathcal{H}} R(h) \leq 2\sqrt{\frac{2\log\frac{1}{\delta}}{m}} + 4\mathfrak{R}_m(\mathcal{G}), \tag{2.100}$$

$$R(\hat{h}_m) - \inf_{h\in\mathcal{H}} R(h) \leq 6\sqrt{\frac{2\log\frac{2}{\delta}}{m}} + 4\hat{\mathfrak{R}}_m(\mathcal{G}). \tag{2.101}$$

*Proof.* From Corollary 2.4.1 and Lemma 2.4.1, we have

$$\mathbb{P}\Big\{\hat{\text{err}}_m(\mathcal{H}) \leq \epsilon + 2\mathfrak{R}_m(\mathcal{G})\Big\} \geq 1 - \exp\Big(-\frac{m\epsilon^2}{2}\Big). \tag{2.102}$$

Let $\delta = \exp\big(-\frac{m\epsilon^2}{2}\big)$. Solving for $\epsilon$ yields

$$\exp\Big(-\frac{m\epsilon^2}{2}\Big) = \delta \tag{2.103}$$

$$-\frac{m\epsilon^2}{2} = \log(\delta) \tag{2.104}$$

$$\epsilon = \sqrt{\frac{2\log\big(\frac{1}{\delta}\big)}{m}}, \tag{2.105}$$

which can then be substituted back into (2.102) to give

$$\hat{\text{err}}_m(\mathcal{H}) \leq \sqrt{\frac{2\log\frac{1}{\delta}}{m}} + 2\mathfrak{R}(\mathcal{G}), \tag{2.106}$$

with probability at least $1-\delta$. Applying Lemma 2.2.1, we get the following high-probability bound on the *estimation error* of ERM:

$$R(\hat{h}_m) - \inf_{h\in\mathcal{H}} R(h) \leq 2\sqrt{\frac{2\log\frac{1}{\delta}}{m}} + 4\mathfrak{R}(\mathcal{G}). \tag{2.107}$$

Additionally, we may use **McDiarmid's inequality** (Theorem 2.3.3) with $f(z_1, ..., z_m) = \hat{\mathfrak{R}}_m(\mathcal{G})$ to obtain a data-dependent upper bound for the *estimation error* of ERM-$\mathcal{H}$. Indeed, we have

$$|f(z_1, ..., z_i, ..., z_m) - f(z_1, ..., z_i', ..., z_m)| \leq \frac{2}{m}, \tag{2.108}$$

for the same reason as in Corollary 2.4.1. Thus,

$$\mathbb{P}\Bigg\{\mathfrak{R}(\mathcal{G}) \leq \hat{\mathfrak{R}}_m(\mathcal{G}) + \sqrt{\frac{2\log\frac{1}{\delta}}{m}}\Bigg\} \geq 1-\delta, \tag{2.109}$$

By letting $\delta = \delta/2$ in both (2.107) and (2.109), and using the union bound, we obtain

$$R(\hat{h}_m) - \inf_{h\in\mathcal{H}} R(h) \leq 6\sqrt{\frac{2\log\frac{2}{\delta}}{m}} + 4\hat{\mathfrak{R}}_m(\mathcal{G}), \tag{2.110}$$

with probability at least $1-\delta$.

$\square$

**Remark 2.4.1.** *In general, computing the Rademacher complexity of a given hypothesis space $\mathcal{H}$ with respect to a binary classification problem $\mathbb{P}$ is not possible since $\mathbb{P}$ is unknown. However, the empirical Rademacher complexity appearing in* (2.101) *depends only on a specific sample and can be estimated [MRT18, p. 38]. In the next section, we will demonstrate how the Rademacher complexity can be upper-bounded by a purely combinatorial notion, which is often easier to deal with: the VC dimension.*

## VC dimension and a fundamental theorem on learnability

In contrast to the concept of Rademacher complexity, which is a localized measure of complexity that is then averaged considering the relative likelihood of different samples to occur, the VC dimension is a globalized, worst-case measure of capacity/complexity. Its definition concerns the behavior of the *growth function* $\Pi(\mathcal{G}, \cdot)$, which we introduce in Definition 2.4.3. Given $m \in \mathbb{N}$, $\Pi(\mathcal{G}, m)$ represents the largest number of distinct classifications realizable by $\mathcal{G}$ over a sample of $m$ points. It is evident that $\Pi(\mathcal{G}, m) \leq 2^m$, and that if $\Pi(\mathcal{G}, m) = 2^m$ for a given $m$, then $\Pi(\mathcal{G}, n) = 2^n$ for every $n \leq m$. As such, it is natural to study $\Pi(\mathcal{G}, m)$ for $m > d$, where $d$ is the largest natural number such that $\Pi(\mathcal{G}, d) = 2^d$. This critical point is precisely the VC dimension (see Definition 2.4.4).

Chronologically, the VC dimension was introduced into machine learning theory roughly 30 years earlier than the Rademacher complexity [VC68]. At the time, researchers where interested in characterizing which classes possessed the uniform convergence property; that is, characterizing for which $\mathcal{H}$ the convergence $\hat{\text{err}}_m(\mathcal{H}) \xrightarrow{\mathbb{P}} 0$ could be shown uniformly over all $\mathbb{P}$. After discussing VC-based upper bounds on the estimation error of ERM-$\mathcal{H}$, we present this fundamental result as Theorem 2.4.4. For a more detailed treatment, readers are referred to the original work by Vapnik.

**Definition 2.4.3** (**Growth function**)**.** *Let $\mathcal{G}$ be a family of functions $g : \mathcal{Z} \to \{0, 1\}$, then the* growth function *of $\mathcal{G}$ is given by*

$$\Pi(\mathcal{G}, m) = \max_{\{z_1,\dots,z_m\}\subset\mathcal{Z}} \#\Big\{\Big(g(z_1), ..., g(z_m)\Big) : g \in \mathcal{G}\Big\}. \tag{2.111}$$

**Example 2.4.1** (**Intervals**)**.** *Let $\mathcal{Z} = \mathbb{R}$ and $\mathcal{G} = \{\mathbf{1}_A : A \subset \mathbb{R}$ is a closed interval $\}$, then $\Pi(\mathcal{G}, m) = \frac{(m+1)m}{2} + 1$. This example can be generalized to produce hypothesis classes based on any family of geometric shapes on $\mathbb{R}^n$.*

*Details.* Consider a fixed sample $z_1 < z_2 < ... < z_m \in \mathbb{R}$. We can partition the line $\mathbb{R} = (-\infty, z_1] \cup (z_1, z_2] \cup ... \cup (z_m, \infty)$ so that the starting point $a$ of any closed interval $[a, b]$ lies exclusively in one of those subpartitions. Notice that the exact place the point lands in each partition doesn't matter for the sake of classification, except only for the case of the trivial classification $(0, 0, ..., 0)$ (e.g., $[z_1, b)$ cannot generate it). Let's start by counting this trivial classification: 1. After we have done this, the classifications generated by each equivalence class are guaranteed to be distinct from each other class. Now, to finish counting the distinct classifications one only has to note that for each class $k \in [m+1]$, there are only $m+1-k$ classifications that can be achieved by moving the end point $b$ futher right. As such, $\Pi(\mathcal{G}, m) = 1 + (m + (m-1) + ... + 1) = \frac{(m+1)m}{2} + 1$, as desired. □

**Remark 2.4.2.** *Observe that any set of functions from $\mathcal{Z} \to \{0, 1\}$ essentially constitutes a collection of indicator functions. Consequently, it can be seamlessly interchanged with the corresponding family of subsets of $\mathcal{Z}$ that these functions represent.*

**Definition 2.4.4 (VC dimension).** *Given a collection $\mathcal{G}$ of functions $g : \mathcal{Z} \to \{0, 1\}$, then the VC-dimension of $\mathcal{G}$ is defined as:*

$$\text{VC-dim}(\mathcal{G}) = \sup\Big\{m : \Pi(\mathcal{G}, m) = 2^m\Big\}, \tag{2.112}$$

*where $\Pi(\mathcal{G}, \cdot)$ is the growth function of $\mathcal{G}$.*

**Remark 2.4.3.** *In words, the VC dimension* $\text{VC-dim}(\mathcal{G})$ *represents the maximum sample size for which there exists a sample that can be classified in every conceivable way by the class $\mathcal{G}$, a condition often referred to as being "shattered".*

**Example 2.4.2 (Intervals).** *The VC dimension of the family in example Example* 2.4.1 *is 2.*

**Example 2.4.3 (Linear classifiers).** *Let $\mathcal{Z} = \mathbb{R}^d$ and $\mathcal{G} = \{z \mapsto \text{sign}(\langle w, z\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$ be the class of linear classifiers. Then,* $\text{VC-dim}(\mathcal{G}) = d + 1$.

*Details.* Consider the set $S = \{0, e_1, ..., e_d\}$, where the vectors $e_1, ..., e_d$ represent the canonical basis of $\mathbb{R}^d$. We show that all possible classifications of $S$ are achievable by $\mathcal{G}$. Indeed, given the classification $S_+ = \{0, e_{i_1}, ..., e_{i_k}\}$ and $S_- = S - S_+$, we can let $b = 0$ and $w$ be $+1$ in the $S_+$ indices and $-1$ in the remaining indices. If $0 \notin S_+$, then we can adjust $b = -0.5$. Such construction covers all $2^{d+1}$ classifications of these $d+1$ points so that $\text{VC-dim}(\mathcal{G}) \geq d + 1$.

The other inequality follows from Theorem 2.4.2 that will be presented ahead. Indeed, suppose $\text{VC-dim}(\mathcal{G}) > d + 1$. Then, there would be $d + 2$ points which could be classified in all possible ways by $\mathcal{G}$. From Theorem 2.4.2, there would exist a labelling $S_+ = \{z_1, ..., z_T\}$ and $S_- = \{z_{T+1}, ..., z_{d+2}\}$ such that the convex hulls intersected non-trivially, i.e., $z \in \text{hull}(S_+) \cap \text{hull}(S_-)$. As such, there would be $p_1, ..., p_{d+2} \geq 0$ such that $\sum_{t=1}^{T} p_t = \sum_{t=T+1}^{d+2} p_t = 1$ and $z = \sum_{t=1}^{T} p_t z_t = \sum_{t=T+1}^{d+2} p_t z_t$. However, this would lead to a contradiction, as $\langle w, z\rangle + b = \sum_{t=1}^{T} p_t(\langle w, z_t\rangle + b) = \sum_{t=T+1}^{d+2} p_t(\langle w, z_t\rangle + b)$, but the last two expressions would have to have different signs. □

**Theorem 2.4.2** (**Radon** [Rad21, p. 113-115])**.** *Let $S \subset \mathbb{R}^d$ be a set of at least $d+2$ elements. There are $S_+$ and $S_-$ such that $S_+ \cup S_- = S$, $S_+ \cap S_- = \emptyset$, but* $\text{hull}(S_+) \cap \text{hull}(S_-) \neq \emptyset$.

**Example 2.4.4** (**Spheres**)**.** *Let $\mathcal{Z} = \mathbb{R}^d$ and $\mathcal{G} = \{z \mapsto \mathbf{1}_{||z-c||_2^2 \leq r} : c \in \mathbb{R}^d\}$ be the class of spherical classifiers. Then,* $\text{VC-dim}(\mathcal{G}) \leq d+2$.

*Details.* Suppose for the sake of contradiction that $\text{VC-dim}(G) > d+2$. There would be a set of $d+3$ points $z_1, ..., z_{d+3} \in \mathbb{R}^d$ shattered by $\mathcal{G}$. In particular, if we considered the maps from $\mathbb{R}^d$ to $\mathbb{R}^{d+1}$ given by $\phi(z) = (z, ||z||_2^2)$ and $\psi(c) = (-2c, 1)$, then there would be a classification of the set $\phi(\{z_1, ..., z_{d+3}\})$ that wouldn't be realizable by linear classifiers in $\mathbb{R}^{d+1}$, as the VC dimension of linear classifiers in $\mathbb{R}^{d+1}$ is $d+2$. Fix the ball $(c, r)$ that realizes this classification for $\{z_1, ..., z_{d+3}\}$ in $\mathbb{R}^d$. If we look at the decision rule:

$$||z-c||_2^2 \leq r \iff ||z||_2^2 - 2\langle z, c\rangle + ||c||_2^2 \leq r \iff \langle \psi(c), \phi(z)\rangle + b \leq 0, \tag{2.113}$$

where $b = ||c||_2^2 - r$, it is evident that there would be a linear classifier in $\mathbb{R}^{d+1}$ that realized the impossible classification for $\phi(\{z_1, ..., z_{d+3}\})$. This is absurd, which implies that $\text{VC-dim}(\mathcal{G}) \leq d+2$. □

Now that we have a clear understanding of the VC dimension and the growth function, we return to our task of bounding $\text{err}_m(\mathcal{H})$. The following lemma due to Massart enables us to remove the dependency on $\mathbb{P}$.

**Lemma 2.4.2** (**Massart** [Mas00])**.** *Let $\mathcal{A} \subset \mathbb{R}^m$ be a finite set and $r = \max_{x\in\mathcal{A}} ||x||_2$, then*

$$\mathbb{E}_\sigma\left[\sup_{x\in\mathcal{A}} \sum_{i=1}^m \sigma_i x_i\right] \leq r\sqrt{2\log(\#\mathcal{A})}, \tag{2.114}$$

*where the expectation is taken with respect to $\sigma_1, ..., \sigma_m \overset{iid}{\sim} \mathbf{Unif}\{-1, +1\}$.*

*Proof.* Given $x \in \mathcal{A}$, note that:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(t\sum_{i=1}^m \sigma_i x_i\right)\right] &= \mathbb{E}\left[\prod_{i=1}^m \exp(t\sigma_i x_i)\right] && (2.115)\\
&= \prod_{i=1}^m \mathbb{E}\Big[\exp(t\sigma x_i)\Big] && (2.116)\\
&\leq \prod_{i=1}^m \exp\left(\frac{t^2|x_i|^2}{2}\right) && (2.117)\\
&= \exp\left(\frac{t^2||x||_2^2}{2}\right) && (2.118)\\
&\leq \exp\left(\frac{t^2 r^2}{2}\right), && (2.119)\\
& && (2.120)
\end{aligned}$$

where the third line follows from **Hoeffding's lemma** (Lemma 2.3.1). To avoid clutter, lets denote $X_j = \sum_{i=1}^m \sigma_i x_i^{(j)}$, where $\mathcal{A} = \{x^{(1)}, ..., x^{(n)}\}$. Then, by **Jensen's inequality**:

$$\begin{aligned}
\exp\left(t\mathbb{E}\Big[\max_{1\le j\le n} X_j\Big]\right) &\le \mathbb{E}\Big[\exp\left(t\max_{1\le j\le n} X_j\right)\Big] && (2.121)\\
&= \mathbb{E}\Big[\max_{1\le j\le n}\Big(\exp(tX_j)\Big)\Big] && (2.122)\\
&\le \mathbb{E}\Big[\sum_{j=1}^n\Big(\exp(tX_j)\Big)\Big] && (2.123)\\
&\le n\exp\left(\frac{t^2r^2}{2}\right). && (2.124)
\end{aligned}$$

By taking the log, we get:

$$\mathbb{E}\Big[\max_{1\le j\le n} X_j\Big] \le \frac{\log(n)}{t} + \frac{tr^2}{2}. \tag{2.125}$$

Minimizing over $t > 0$ yields:

$$\begin{aligned}
0 &= -\frac{\log(n)}{t^2} + \frac{r^2}{2} && (2.126)\\
2\log(n) &= r^2t^2 && (2.127)\\
\frac{\sqrt{2\log(n)}}{r} &= t. && (2.128)
\end{aligned}$$

Substituting $t = \frac{\sqrt{2\log(n)}}{r}$ back into the previous inequality finishes the proof. $\square$

**Remark 2.4.4.** *If we symmetrize the set $\mathcal{A}$ by letting $\mathcal{A} = \mathcal{A} \cup -\mathcal{A}$, then* **Massart's lemma** *(Lemma 2.4.2) gives the following*

$$\mathbb{E}_\sigma\left[\sup_{x\in\mathcal{A}}\left|\sum_{i=1}^m \sigma_i x_i\right|\right] \le r\sqrt{2\log(2\#\mathcal{A})}. \tag{2.129}$$

Now, by chaining Symmetrization (Lemma 2.4.1) and Massart's lemma (Lemma 2.4.2), we get a distribution-free bound on $\mathrm{err}_m(\mathcal{H})$.

**Corollary 2.4.2.** *Let $\mathcal{G}$ be a family of functions $g : \mathcal{Z} \to \{0, 1\}$, then*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^m\Big(g(z_i) - \mathbb{E}_{z\sim\mathbb{P}}[g(z)]\Big)\right|\right] \le 2\sqrt{\frac{2\log(2\Pi(\mathcal{G},m))}{m}}, \tag{2.130}$$

*where $\Pi(\mathcal{G},\cdot)$ is the growth function of $\mathcal{G}$.*

*Proof.* By applying the two lemmas, we get

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^{m}\Big(g(z_i)-\mathbb{E}_{z\sim\mathbb{P}}[g(z)]\Big)\right|\right] \le \frac{2}{m}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{m}\sigma_i g(z_i)\right|\right] \tag{2.131}$$

$$\le \frac{2\sqrt{m}}{m}\sup_{\{z_1,\ldots,z_m\}\subset\mathcal{Z}}\sqrt{2\log(2\#\mathcal{A})} \tag{2.132}$$

$$\le 2\sqrt{\frac{2\log(2\Pi(\mathcal{G},m))}{m}}, \tag{2.133}$$

where $\mathcal{A}=\Big\{\Big(g(z_1),\ldots,g(z_m)\Big):g\in\mathcal{G}\Big\}$. □

Before applying Corollary 2.4.2 to bound the estimation error of ERM-$\mathcal{H}$, we note the following fundamental result on the growth function (Lemma 2.4.3), which states that $\Pi(\mathcal{G},m)$ grows polynomially for $m \geq d = \text{VC-dim}(\mathcal{G})$.

**Lemma 2.4.3** (**Sauer-Shelah** [Sau72])**.** *Let $\mathcal{G}$ be a collection of functions $g:\mathcal{Z}\to\{0,1\}$ with* VC-dim$(\mathcal{G})=d$*. Then, for all $m\in\mathbb{N}$, the following holds:*

$$\Pi(\mathcal{G},m) \le \sum_{i=0}^{d}\binom{m}{i}. \tag{2.134}$$

**Corollary 2.4.3.** *Let $\mathcal{G}$ be a collection of functions $g:\mathcal{Z}\to\{0,1\}$ with* VC-dim$(\mathcal{G})=d$*, then, for all $m\geq d$,*

$$\Pi(\mathcal{G},m) \le \left(\frac{em}{d}\right)^{d}. \tag{2.135}$$

*Proof.* For $m \geq d$,

$$\Pi(\mathcal{G},m) \le \sum_{i=0}^{d}\binom{m}{i} \tag{2.136}$$

$$\le \sum_{i=0}^{m}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} \tag{2.137}$$

$$= \left(\frac{m}{d}\right)^{d}\sum_{i=0}^{m}\binom{m}{i}\left(\frac{d}{m}\right)^{i} \tag{2.138}$$

$$= \left(\frac{m}{d}\right)^{d}\left(1+\frac{d}{m}\right)^{m}, \tag{2.139}$$

$$\le \left(\frac{em}{d}\right)^{m} \tag{2.140}$$

where (2.136) follows from **Sauer-Shelah's lemma** (Lemma 2.4.3), (2.139) follows from the **Binomial theorem**, and (2.140) follows from the well-known inequality $(1-x)\leq e^{-x}$. □

Equipped with the closed-form expression of Corollary 2.4.3, we now present the definitive VC bound on the estimation error of ERM-$\mathcal{H}$ (see Theorem 2.4.3).

**Theorem 2.4.3.** *Let $\mathcal{H}$ be a hypothesis class and $\mathcal{G} = \{(x, y) \mapsto \mathbf{1}_{h(x)\neq y} : h \in \mathcal{H}\}$. If $d = \text{VC-dim}(\mathcal{G})$, then, with probability at least $1 - \delta$, we have the following bound on the estimation error of ERM-$\mathcal{H}$:*

$$R(\hat{h}_m) - \inf_{h\in\mathcal{H}} R(h) \leq 2\sqrt{\frac{2\log\frac{1}{\delta}}{m}} + 8\sqrt{\frac{2\log(2) + 2d\log(\frac{em}{d})}{m}}. \tag{2.141}$$

*Proof.* From Corollary 2.4.1 and Corollary 2.4.2, we have the following bound

$$\hat{\text{err}}_m(\mathcal{H}) \leq \sqrt{\frac{2\log\frac{1}{\delta}}{m}} + 4\sqrt{\frac{2\log(2\Pi(\mathcal{G}, m))}{m}}, \tag{2.142}$$

with probability at least $1 - \delta$. Then, by Corollary 2.4.3,

$$\hat{\text{err}}_m(\mathcal{H}) \leq \sqrt{\frac{2\log\frac{1}{\delta}}{m}} + 4\sqrt{\frac{2\log(2) + 2d\log(\frac{em}{d})}{m}}, \tag{2.143}$$

Applying Lemma 2.2.1 finishes the proof:

$$R(\hat{h}_m) - \inf_{h\in\mathcal{H}} R(h) \leq 2\sqrt{\frac{2\log\frac{1}{\delta}}{m}} + 8\sqrt{\frac{2\log(2) + 2d\log(\frac{em}{d})}{m}}. \tag{2.144}$$

$\square$

At last, the fundamental theorem of statistical learning theory (Theorem 2.4.4).

**Theorem 2.4.4** (**Vapnik-Chervonenkis** [VC68]). *Let $\mathcal{G}$ be a collection of indicator functions $g : \mathcal{Z} \to \{0, 1\}$ with $\text{VC-dim}(\mathcal{G}) = d$, then*

*i. If $d < +\infty$, we have*

$$\sup_{\mathbb{P} \text{ measure over } \mathcal{Z}} \mathbb{E}\left[\sup_{g\in\mathcal{G}} \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathbb{E}[g(z)]\right|\right] = \mathcal{O}\left(\sqrt{\frac{\log m}{m}}\right). \tag{2.145}$$

*ii. Otherwise, if $d = \infty$,*

$$\limsup_{m\to\infty} \sup_{\mathbb{P} \text{ measure over } \mathcal{Z}} \mathbb{E}\left[\sup_{g\in\mathcal{G}} \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathbb{E}[g(z)]\right|\right] \geq e^{-1} > 0. \tag{2.146}$$

*Proof.* We proceed by cases.

i. By **Corollaries 2.4.2 and 2.4.3**, we have that

$$\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}} \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathbb{E}[g(z)]\right|\right] &\leq 2\sqrt{\frac{2\log(2\Pi(\mathcal{G}, m))}{m}} &(2.147)\\
&\leq 2\sqrt{\frac{2d\log\left(\frac{2^{\frac{1}{d}}em}{d}\right)}{m}} &(2.148)\\
&= \mathcal{O}\left(\sqrt{\frac{\log m}{m}}\right). &(2.149)
\end{aligned}$$

Since the rhs doesn't depend on $\mathbb{P}$, we can safely take the supremum without affecting the rhs.

ii. For $m \geq 1$, let $B_m \subset \mathcal{Z}$ be a set of $m$ elements that is shattered by $\mathcal{G}$ and $\mathbb{P}_m = \mathbf{Unif}(B_m)$. Then,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathbb{E}[g(z)]\right|\right] \geq \left(1-\frac{1}{m}\right)^m, \tag{2.150}$$

as we shall soon see. Indeed, let $z_1, ..., z_m \overset{iid}{\sim} \mathbb{P}_m$ and $\hat{B}_m = \left\{b \in B_m : \exists i \in [m] : z_i = b\right\}$. Now, since $B_m$ is shattered by $\mathcal{G}$, we have that there is $\hat{g}_m \in \mathcal{G}$ such that $\hat{g}_m = \mathbf{1}_{\hat{B}_m}$ in $B_m$. Thus,

$$\begin{aligned}
\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathbb{E}[g(z)]\right| &\geq \frac{1}{m}\sum_{i=1}^{m} \hat{g}_m(z_i) - \frac{\#\hat{B}_m}{m} && (2.151)\\
&= 1 - \frac{\#\hat{B}_m}{m} && (2.152)\\
&= \frac{\#\left\{b \in B_m : \forall i \in [m] : z_i \neq b\right\}}{m}. && (2.153)\\
& && (2.154)
\end{aligned}$$

We then take the expected value

$$\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathbb{E}[g(z)]\right|\right] &\geq \mathbb{E}\left[\frac{\#\left\{b \in B_m : \forall i \in [m] : z_i \neq b\right\}}{m}\right] && (2.155)\\
&= \frac{1}{m}\sum_{b\in B_m} \mathbb{P}(\forall i \in [m] : b \neq z_i) && (2.156)\\
&= \left(1-\frac{1}{m}\right)^m. && (2.157)
\end{aligned}$$

By taking the sup with respect to $\mathbb{P}$, we can only increase the lhs so that

$$\sup_{\mathbb{P} \text{ measure over } \mathcal{Z}} \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathbb{E}[g(z)]\right|\right] \geq \left(1-\frac{1}{m}\right)^m. \tag{2.158}$$

Finally, the result follows by taking the limsup.

□

# 3 The learning problem

The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work.

*John von Neumann*

In this chapter, we specialize the learning-theoretic discussion of Chapter 2 to the malignancy classification problem. We begin by recapitulating the standard model of binary classification, as introduced in Section 2.1. Let $\mathbb{P}$ be a probability distribution over $\mathcal{X} \times \{0,1\}$ and $(x_1, y_1), ..., (x_m, y_m) \overset{iid}{\sim} \mathbb{P}$ be a sample of size $m$, then the goal of binary classification is to learn from the sample a measurable function $\hat{h}_m$ such that

$$\mathbb{P}(\hat{h}_m(x) \neq y) \text{ is small.} \tag{3.1}$$

To design a good diagnostic system for our problem, however, accuracy is not enough. This is neatly summarized by the following remark:

> *Any model suitable for clinical use must be guaranteed to avoid undetected cancers up to a tolerable margin of uncertainty, while still reducing the number of unnecessary biopsies.*

To formalize these requirements, in Section 3.1, we introduce the notions of predictive values (ppv/precision and npv), sensitivity (recall), and specificity, which serve as more suitable performance metrics for the design of reliable malignancy classification systems. To address the absence of a cost model for classification errors and to facilitate discussions regarding predictive values, in Section 3.2 we introduce scoring functions and threshold-based classifiers. Finally, in Section 3.3, we review some learning-theoretic results of Vemuri and Srebro [VS20] on the generalization of the empirical predictive value curves that we later employ for validation purposes. It's worth noting that throughout the chapter, the positive class is used to represent benign lesions.

## 3.1 Classification metrics for malignancy prediction

There are numerous metrics for evaluating the performance of a binary classifier on a sample [Tha20], each with their particular use cases. The medical community around diagnostic tests has settled with the usage of predictive values, sensitivity, and specificity [Tre17], all of which can be derived from the confusion matrix (see Definition 3.1.1). In this section, we review these four metrics and discuss how they can be used to formalize the requirements laid out at the beginning of the chapter.

**Definition 3.1.1** (**Confusion matrix**). *Given a sample* $(x_1, y_1), ..., (x_m, y_m)$ *and a classifier* $h$*, then the confusion matrix of* $h$ *with respect to the sample is given by*

| | *Predicted Negative* | *Predicted Positive* |
|---|---|---|
| *Actual Negative* | TN | FP |
| *Actual Positive* | FN | TP |

*where* $\mathrm{TN}, \mathrm{TP}, \mathrm{FN}, \mathrm{FP}$ *are, respectively, the total number of true negatives, true positives, false negatives, and false positives obtained by comparing* $y_1, ..., y_m$ *with* $h(x_1), ..., h(x_m)$*.*

With the confusion matrix, we can now define the predictive values (Definition 3.1.2), sensitivity, and specificity (Definition 3.1.3).

**Definition 3.1.2** (**Empirical predictive values**). *Given a sample* $(x_1, y_1), ..., (x_m, y_m)$ *and a classifier* $h$ *such that its confusion matrix with respect to the sample is* $(\mathrm{TN}, \mathrm{FP}, \mathrm{FN}, \mathrm{TP})$*, then its empirical positive and negative predictive values are given by*

$$\mathrm{p\hat{p}v}(h) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{3.2}$$

$$\mathrm{n\hat{p}v}(h) = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}. \tag{3.3}$$

**Remark 3.1.1.** *The predictive values capture the notion of confidence in a classifier's prediction. For instance, consider a model* $h$ *with* $\mathrm{p\hat{p}v}(h) \geq 1 - \epsilon$*. In the context of malignancy prediction, such a model is virtually certain, with a margin of uncertainty of* $\epsilon$*, to be correct when asserting that a patient does not have a malignant lesion. This accreditation underscores the model's high reliability in preventing cases of cancers from going undetected.*

**Definition 3.1.3** (**Sensitivity/Specificity**). *Given a sample* $(x_1, y_1), ..., (x_m, y_m)$ *and a classifier* $h$ *such that its confusion matrix with respect to the sample is* $(\mathrm{TN}, \mathrm{FP}, \mathrm{FN}, \mathrm{TP})$*, then its sensitivity and specificity with respect to the sample are given by*

$$\text{sensitivity}(h) = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3.4}$$

$$\text{specificity}(h) = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{3.5}$$

**Remark 3.1.2.** *Sensitivity and specificity express how much of the positive and negative instances of a sample the classifier was able to discover. In the context of malignancy prediction, a classifier $h$ with high sensitivity is able to discover most of the benign lesions, and as such is able to minimize the number of unnecessary biopsies.*

Together the notions of predictive values, sensitivity, and specificity fully specify the desired performance metrics necessary for the design of a good diagnostic system. Concretely, a good system is one that maximizes sensitivity without compromising the positive predictive value. In other words, high positive predictive value is top priority. Only after achieving a minimum positive predictive value level one switches to maximizing sensitivity.

**Remark 3.1.3 (Precision/Recall).** *We note that, by inverting the labels so that the positive class refers to benign lesions, we are able to use the lighter notation of precision (to mean the positive predictive value) and recall (to mean sensitivity), which we borrow from the information retrieval community. Thus, our goal becomes to maximize recall without compromising precision. The advantages of this change will become clear once we start relying on the Area Under the Precision-Recall Curve (AUPRC) for model selection and validation.*

## 3.2 Scoring functions and threshold-based classifiers

In Section 3.1, we did not require a specific values for the classification metrics discussed. Instead, we vaguely referred to a good diagnostic system as one with high precision and recall which somehow prioritizes precision over recall. This degree of freedom is left on purpose and motivates a natural approach involving scoring functions and threshold-based classifiers. See the definition below.

**Definition 3.2.1 (Threshold-based classifiers).** *Given a scoring function $f : \mathcal{X} \to [0, 1]$, we define its family of threshold-based classifiers to be $\{h_{f,t}\}_{t\in[0,1]}$, where $h_{f,t}$ is given by*

$$h_{f,t}(x) := \mathbf{1}_{f(x)>t}. \tag{3.6}$$

With scoring functions, the classification metrics of the last section become functions of the threshold $t$. See definitions 3.2.2 and 3.2.3. Additionally, we are able to define the precision-recall curve (Definition 3.2.4).

**Definition 3.2.2** (**Empirical predictive value curves**). *Let $f : \mathcal{X} \to [0,1]$ be a scoring function and $t \in [0,1]$ be a threshold level, then its empirical predictive value curves are given by*

$$\mathrm{p\hat{p}v}(f,t) = \mathrm{p\hat{p}v}(h_{f,t}), \tag{3.7}$$
$$\mathrm{n\hat{p}v}(f,t) = \mathrm{n\hat{p}v}(h_{f,t}). \tag{3.8}$$

**Definition 3.2.3** (**Sensitivity/Specificity curves**). *Let $f : \mathcal{X} \to [0,1]$ be a scoring function and $t \in [0,1]$ be a threshold level, then its sensitivity/specificity curves are given by*

$$\mathrm{sensitivity}(f,t) = \mathrm{sensitivity}(h_{f,t}), \tag{3.9}$$
$$\mathrm{specificity}(f,t) = \mathrm{specificity}(h_{f,t}). \tag{3.10}$$

**Definition 3.2.4** (**Precision-recall curve**). *Let $f : \mathcal{X} \to [0,1]$ be a scoring function, then its precision-recall curve is obtained by plotting precision as a function of recall for increasing threshold levels $t \in [0,1]$. For an example, consider the right column of Figure 2.*

One advantage of such approach is that it subdivides the binary classification problem into two subproblems: to choose a data-dependent scoring function $\hat{f}_m \in \mathcal{F}$ and to select a threshold level $t \in [0,1]$. This is preferable as it allows one to perform modeling experiments without having to commit to any particular trade-off between the classification metrics [Har19; Hon+22]. Concretely, this enables the decision component of the analysis to be performed at a later moment, preferably when more detailed information about the usage of the model, such as a cost model for the classification mistakes, becomes available.

Another motivating factor for the introduction of scoring functions and threshold-based classifiers lies in the empirical predictive values. In particular, to analyze the generalization of the empirical predictive values one inevitably has to deal with the denominators

$$\mathrm{PP} = \mathrm{TP} + \mathrm{FP} \quad \text{and} \quad \mathrm{PN} = \mathrm{TN} + \mathrm{FN}, \tag{3.11}$$

which constitute the number of points the classifier predicted as positive (PP) and negative (PN) for a given sample. The determination of PP and PN for an arbitrary classifier $h$ and sample $S$ isn't trivial, but it can be easily done in the case of scoring functions by choosing thresholds based on the empirical quantile function. Indeed, let $f$ be a scoring function and $(x_1, y_1), ..., (x_m, y_m)$ be a sample so that $f_1 \leq f_2 \leq ... \leq f_m$ is $f(x_1), f(x_2), ..., f(x_m)$ in ascending order. Moreover, let $k \in [m-1]$ and $\alpha_k = k/m$ so that the empirical quantile function $\hat{q}_f$ is given by:

$$\hat{q}_f(\alpha_k) = \frac{f_{m-k} + f_{m-k+1}}{2}, \quad \text{for } k = 1, ..., m-1. \tag{3.12}$$

Then, the empirical predictive values simplify to:

$$\hat{\text{ppv}}(f, \hat{q}_f(\alpha_k)) = \frac{1}{k} \sum_{i=1}^{m} \mathbf{1}_{y_i=1 \wedge f(x_i) > \hat{q}_f(\alpha_k)}, \tag{3.13}$$

$$\hat{\text{npv}}(f, \hat{q}_f(\alpha_k)) = \frac{1}{m-k} \sum_{i=1}^{m} \mathbf{1}_{y_i=0 \wedge f(x_i) \leq \hat{q}_f(\alpha_k)}, \tag{3.14}$$

which suggests a natural reparameterization by considering the threshold levels to be given by the *positive rate* $\alpha_k$:

$$\hat{\text{ppv}}(f, \alpha_k) = \frac{1}{k} \sum_{i=1}^{m} \mathbf{1}_{y_i=1 \wedge f(x_i) > \hat{q}_f(\alpha_k)}, \tag{3.15}$$

$$\hat{\text{npv}}(f, \alpha_k) = \frac{1}{m-k} \sum_{i=1}^{m} \mathbf{1}_{y_i=0 \wedge f(x_i) \leq \hat{q}_f(\alpha_k)}, \tag{3.16}$$

$$\tag{3.17}$$

**Remark 3.2.1.** *Note that, under the framework of learning theory, the expressions in equations* (3.15) *and* (3.16) *are almost sums of i.i.d. random variables, except for the dependence introduced by the empirical quantile function.*

Given the previous discussion on scoring functions and their benefits to our approach, in this thesis we shall consider the malignancy prediction problem as the problem of selecting a data-dependent scoring function $\hat{f}_m$ which possesses good classification curves, as introduced in definitions 3.2.2, 3.2.3, and 3.2.4. In particular, we will focus on obtaining (and establishing statistical control over) positive predictive value curves which are as close as possible to 1 in the critical interval $(0, 0.5]$ of positive rate/threshold levels $\alpha_k$. Moreover, as a secondary goal, we will pursue the maximization of the AUPRC.

## 3.3 Predictive value generalization bounds

Having discussed the relevant classification metrics in Sections 3.1 and 3.2, in this section we review the learning-theoretic results of Vemuri and Srebro [VS20] on the generalization of predictive value curves. We begin by introducing the population counterparts of the empirical quantile function and predictive values. Let $k \in [m-1]$ and $\alpha_k = k/m$, then

$$q_f(\alpha_k) = \sup\{t : \mathbb{P}_{\mathcal{X}}\{x : f(x) > t\} = \alpha_k\}, \tag{3.18}$$

$$\text{ppv}(f, \alpha_k) = \mathbb{P}(y = 1 \mid f(x) > q_f(\alpha_k)), \tag{3.19}$$

$$\text{npv}(f, \alpha_k) = \mathbb{P}(y = 0 \mid f(x) \leq q_f(\alpha_k)). \tag{3.20}$$

A natural question is how $\hat{\text{ppv}}$ and $\hat{\text{npv}}$ relate to ppv and npv. Lemma 3.3.1 shows that the empirical predictive values are almost unbiased estimators of their population counterparts.

**Lemma 3.3.1** (**Bias of empirical predictive values** [VS20, Lemma 1])**.** *For any $k = 1, 2, ..., m-1$ we have that*

$$\left|\mathbb{E}\left[\hat{\text{ppv}}(f, \alpha_k)\right] - \text{ppv}(f, \alpha_k)\right| \leq \frac{m}{2k}\sqrt{\frac{\pi}{2(m-1)}}, \tag{3.21}$$

$$\left|\mathbb{E}\left[\hat{\text{npv}}(f, \alpha_k)\right] - \text{npv}(f, \alpha_k)\right| \leq \frac{m}{2(m-k)}\sqrt{\frac{\pi}{2(m-1)}}. \tag{3.22}$$

**Remark 3.3.1.** *Notice the trade-off between the two bounds. The bias of the positive (negative) predictive value worsens (improves) as $k \to 0$, while it improves (worsens) as $k \to m-1$. If $k$ is away from both extremes, then asymptotically, both empirical predictive values approach their population counterparts. This trade-off is present in all generalization bounds of this section.*

Lemma 3.3.1 together with an application of McDiarmid's inequality (Theorem 2.3.3) to the empirical predictive values allows us to provide confidence bands for the predictive value curves of a fixed scorer $f : \mathcal{X} \to [0, 1]$. See Theorem 3.3.1.

**Theorem 3.3.1** (**Large deviation bound** [VS20, Theorem 2])**.** *With probability at least $1 - \delta$, for all $k = 1, 2, .., m-1$,*

$$\left|\hat{\text{ppv}}(f, \alpha_k) - \text{ppv}(f, \alpha_k)\right| \leq \frac{1}{k}\sqrt{\frac{m \log(\frac{4m}{\delta})}{2}} + \frac{m}{2k}\sqrt{\frac{\pi}{2(m-1)}}, \tag{3.23}$$

$$\left|\hat{\text{npv}}(f, \alpha_k) - \text{npv}(f, \alpha_k)\right| \leq \frac{1}{m-k}\sqrt{\frac{m \log(\frac{4m}{\delta})}{2}} + \frac{m}{2(m-k)}\sqrt{\frac{\pi}{2(m-1)}}. \tag{3.24}$$

*Proof.* Let $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ and $F : \mathcal{Z}^m \to \mathbb{R}$ be defined by $F(z_1, ..., z_m) = \hat{\text{ppv}}(f, \alpha_k)$. We begin by showing that $F$ satisfies the bounded differences condition; i.e., that

$$|F(z_1, ..., z_i, ..., z_m) - F(z_1, ..., z_i', ..., z_m)| \leq \frac{1}{k}, \tag{3.25}$$

for all $z_1, ..., z_i, z_i', ..., z_m \in \mathcal{Z}$ and $i \in [m]$. This is more easily seen geometrically than algebraically. Indeed, recall that $\hat{\text{ppv}}(f, \alpha_k)$ is computed by first sorting $f(x_1), ..., f(x_m)$ and then selecting a threshold $\hat{q}_f(\alpha_k)$ such that the top $k$ scores are above it and the other $m-k$ scores are below it. The $\hat{\text{ppv}}(f, \alpha_k)$ is then the number of points $x_j$ such that $f(x_j)$ lies among the top $k$ scores and $y_j$ is positive divided by $k$. It is clear that perturbing the sample by one point won't alter the denominator $\frac{1}{k}$, so that

$$|F(z_1, ..., z_i, ..., z_m) - F(z_1, ..., z_i', ..., z_m)| \tag{3.26}$$

is $\frac{1}{k}$ times the difference between the counts of true positives across the two sets of top-$k$ scores. As the perturbation can only cause the set $\{x_j : f(x_j) > \hat{q}_f(\alpha_k)\}$ to change by at most one point, the set $\{y_j : f(x_j) > \hat{q}_f(\alpha_k)\}$ can also only change by at most one value. The difference in counts is thus bounded by a difference of values of $y$: $|\mathbf{1}_{y_i=1} - \mathbf{1}_{y_l=1}|$, for some $l \in [m]$. As such, the difference in the positive predictive values cannot be larger than $\frac{1}{k}$, as desired.

This can be similarly shown for $G(z_1, ..., z_m) = \hat{\text{npv}}(f, \alpha_k)$, but with $c_i = \frac{1}{m-k}$ instead of $\frac{1}{k}$. Thus, **McDiarmid's inequality** (Theorem 2.3.3) applies, and we have that, for each $k \in [m-1]$,

$$\mathbb{P}\left\{\left|\hat{\text{ppv}}(f, \alpha_k) - \mathbb{E}[\hat{\text{ppv}}(f, \alpha_k)]\right| > \epsilon\right\} \leq 2\exp\left(\frac{-2k^2\epsilon^2}{m}\right), \tag{3.27}$$

which can be rewritten in the form of a high-probability bound: with probability at least $1 - \delta/2m$,

$$\left|\hat{\text{ppv}}(f, \alpha_k) - \mathbb{E}[\hat{\text{ppv}}(f, \alpha_k)]\right| \leq \frac{1}{k}\sqrt{\frac{m\log(\frac{4m}{\delta})}{2}}. \tag{3.28}$$

A similar analysis yields: with probability at least $1 - \delta/2m$,

$$\left|\hat{\text{npv}}(f, \alpha_k) - \mathbb{E}[\hat{\text{npv}}(f, \alpha_k)]\right| \leq \frac{1}{m-k}\sqrt{\frac{m\log(\frac{4m}{\delta})}{2}}. \tag{3.29}$$

Now, by applying Lemma 3.3.1 and the union bound, we get the desired result. □

Similar to our treatment of the generalization error of ERM-$\mathcal{H}$ via uniform convergence in Section 2.4, Vemuri and Srebro also provide generalization bounds that hold uniformly over a class of scoring functions $\mathcal{F}$. Concretely, this enables one to analyze the generalization performance of a data-dependent choice $\hat{f}_m \in \mathcal{F}$ from its training sample to the population. Before presenting the main result, we introduce two specialized notions of model complexity. See definitions 3.3.1 and 3.3.2.

**Definition 3.3.1** (**Order coefficient**). *Let $\mathcal{F}$ be a family of scoring functions, then its $(m, k)$-order coefficient is given by*

$$\Theta(\mathcal{F}, m, k) = \max_{x \in \mathcal{X}^m} \#\left\{\phi_k(f(x)) : f \in \mathcal{F}\right\}, \tag{3.30}$$

*where $\phi_k : \mathbb{R}^m \to \mathcal{P}([n])$ takes a vector to the set of indices of its top $k$ entries. Ties are resolved using the original order.*

**Definition 3.3.2** (**VC subgraph dimension**). *Let $\mathcal{F}$ be a family of scoring functions, then its VC subgraph dimension is given by*

$$\text{VC-sub}(\mathcal{F}) = \text{VC-dim}(\mathbf{1}_{S(f)}), \tag{3.31}$$

*where $S(f) = \{(x, t) \in \mathcal{X} \times \mathbb{R} : t < f(x)\}$ is the subgraph of f.*

**Remark 3.3.2.** *Vemuri and Srebro show that the* $(m, k)$*-order coefficient can be bounded using the VC subgraph dimension via an argument analogous to Corollary* 2.4.3*. In particular, they show that*

$$\Theta(\mathcal{F}, m, k) \leq \left(\frac{em}{d}\right)^d, \tag{3.32}$$

*where* $d = \text{VC-sub}(\mathcal{F}) < \infty$ *and* $m \geq d$*.*

We are now ready to state the uniform convergence bound.

**Theorem 3.3.2** (**Uniform convergence bound** [VS20, Theorem 3])**.** *With probability at least* $1 - \delta$*, for* $k = 1, ..., m-1$ *and for all* $f \in \mathcal{F}$ *with* $d = \text{VC-dim}(\mathcal{F}) < \infty$*,*

$$\left|\text{p}\hat{\text{p}}\text{v}(f, \alpha_k) - \text{ppv}(f, \alpha_k)\right| \leq \frac{1}{k}\sqrt{2m \cdot \log(8m \cdot \Theta(\mathcal{F}, m, k)^2/\delta)} + \frac{m}{2k}\sqrt{\frac{\pi}{2(m-1)}}, \tag{3.33}$$

$$\left|\text{n}\hat{\text{p}}\text{v}(f, \alpha_k) - \text{npv}(f, \alpha_k)\right| \leq \frac{1}{m-k}\sqrt{2m \cdot \log(8m \cdot \Theta(\mathcal{F}, m, m-k)^2/\delta)} + \frac{m}{2(m-k)}\sqrt{\frac{\pi}{2(m-1)}}. \tag{3.34}$$

# 4 Machine learning experiments

> The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data
>
> *John Tukey*

In this chapter, we explore the development of a machine learning model for addressing the malignancy prediction problem as described in Chapter 3. To facilitate our discussion, we break-up the modeling process into three phases as outlined in the comprehensive scoping review "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare" by Hond et al. [Hon+22]:

1. **Preparation, collection, and checking of the data.** This phase consists of data acquisition and preprocessing. There are many important considerations regarding this process, such as data quality, representativeness, and dataset size. For more details, see [Hon+22, Phase 1: Preparation, collection, and checking of the data] and references therein.

2. **Development of a machine learning model.** This phase is comprised of model selection[1], training, and internal validation. Here, important considerations are over-fitting, algorithmic bias, and transparency of the modeling process. For more details, consult [Hon+22, Phase 2: Development of the AIPM] and references therein.

3. **Validation of the machine learning model.** This phase consists in a comprehensive assessment of generalization, including an external validation experiment conducted by an independent research group. For more details, refer to [Hon+22, Phase 3: Validation of the AIPM] and references therein.

The first phase was carried out by I. Buzzato in collaboration with four breast cancer reference centers in Brazil [Buz+23]. Detailed information concerning data acquisition procedures, inclusion/exclusion criteria, and feature selection [GE03] may be found in the methods section of the paper. The resulting dataset is discussed in Section 4.1.

The second phase is the main concern of our efforts and is described in full detail in Sections 4.2, 4.3, and 4.4. In particular, in Section 4.2, we justify our gradient boosting approach and compare it to other alternatives, describe training and validation

[1] In the context of the scoping review, model selection refers to choosing a modeling approach.

procedures, and detail the computational environment used. In Section 4.3, we report the results of our machine learning experiments. Finally, in Section 4.4, we interpret and discuss the results of the experiments.

Phase three has not been incorporated into this work. At the time of the writing this thesis, we have not secured an independent group for external validation of our experiments. Therefore, we consider this aspect as a potential avenue for future works and discuss it in Chapter 5.

## 4.1 The dataset

The dataset of Buzatto et al. [Buz+23] is comprised of $m = 1929$ data points, divided into training (1236), validation (290), and testing (403). Each data point represents a lesion of size at most 30mm from a patient aged at least 18 years old related to pathologies originating primarily from the breast. Each patient underwent either percutaneous core needle biopsy and/or excisional biopsy [Kum+22], from which the dependent variable (result) was determined by means of a pathological analysis [AIS18]. The training and test datasets are comprised of BI-RADS 3, 4, 5, and 6 lesions, while the validation dataset only contains BI-RADS 3 and 4 lesions. The dataset's features are described in Table 1. For a more thorough explanation of the features, consult the PhD thesis of Buzatto [Buz24].

In this thesis, we introduce minor modifications to the dataset. Specifically, we utilize the complete dataset without splitting it. Furthermore, we invert the labels so that the positive class represents benign lesions. As discussed in Remark 3.1.3, this adjustment enables us to define the project's goals in terms of the precision-recall curve.

## 4.2 Methods

In this section we discuss two learning experiments. The first relies on Theorems 3.3.1 and 3.3.2 for model validation, whereas the second relies on repeated 5-fold cross-validation (see Appendix A.1). Both experiments are based on gradient-boosted regression trees and so we begin by motivating this decision.

### The choice of learning algorithm

There were many available methodologies for this project, including logistic regression [Cra03], deep learning with convolutional neural networks and image transformers [Dos+20; LeC+89], support vector machines [CV95], and tree-based methods such as random forests [Bre01] and gradient-boosted trees [Fri01].

| Features | | |
|---|---|---|
| Name | Data Type | Description |
| Age | Integer | Age in years |
| Size | Float | Size of the biggest lesion |
| Palpable | Binary | 0: not palpable<br>1: palpable |
| Vessels | Binary | 0: not present<br>1: present |
| RI | Float | Resistance index determined by Doppler spectral analysis |
| Shape | Ordinal | 0: oval or round<br>1: irregular |
| Margins | Ordinal | 0: circumscribed<br>1: microlob./indistinct/angular<br>2: spiculated |
| Orientation | Ordinal | 0: parallel<br>1: not parallel |
| **Result** | **Binary** | **0: malignant**<br>**1: benign** |

Table 1 – The dataset's features.

Given adequate resources, each of these strategies could be made to work well. Nonetheless, specific requirements of our problem and dataset made gradient-boosted trees the most compelling choice. In particular, computer vision approaches based on deep learning can be highly effective for pattern recognition [Ben16], as images contain much more information than a handset of tabular features. However, such approaches typically require enormous datasets to work, which clearly ruled them out for our small dataset[2]. To illustrate this point, nVidia recently provided a deep learning-based solution to the problem we are dealing with and their data set comprised roughly 8 million images of breast ultrasound exams [She+21].

Support vector machine-based modeling[3] is perhaps the closest there is to what we have discussed in Chapter 2. Unfortunately, support vector machines typically require a great deal of feature engineering through kernel design to be effective [SS18]. Furthermore, they do not inherently provide probability estimates [Pla+99], which led to their dismissal for our current purposes[4].

Other methods, including random forests, gradient boosting, and logistic regression, were already evaluated on this dataset, resulting in similar performance across all 9 method choices [Buz+23]. Given this, we chose to focus exclusively on gradient boosting

[2] Less than half of the lesions that comprise our dataset have associated US images. For more information, see [Buz24].

[3] Note that Vladimir Vapnik was involved in the development of both statistical learning theory and support vector machines.

[4] This requirement is discussed in [Hon+22].

in this thesis, as the method is widely recognized as being the go-to method for tabular prediction tasks[5] [Car23; GOV22; Nie16] and possesses many good properties such as robustness against uninformative features [HTF09, Chapter 10].

## Learning-theoretic experiment

To evaluate the practical utility of the predictive value generalization bounds for model validation, we experimented with two sets of models. The first set comprised XGBoost models trained on different train-test splits: $m = 965$, $m = 643$, and $m = 386$, corresponding to one-half, one-third, and one-fifth of the full dataset for validation. The model's hyperparameters were tuned on the basis of a grid search procedure employing 10 repetitions of 5-fold cross-validation (see Table 2 for the grid used). For each split, the associated model was refit on the whole training set ($m = 964$, $m = 1286$, and $m = 1543$) using the hyperparameters found by cross-validation. The refitted models were then evaluated on the corresponding test sets by recording their predictive value and precision-recall curves. To obtain generalization bands for the predictive values, we applied Theorem 3.3.1 with $\delta = 0.05$.

| Hyperparameter | Values |
|---|---|
| Learning rate | 0.3, 0.1, 0.01 |
| Boosting rounds | 10, 100, 1000 |
| Maximum depth | 4 |
| Minimum child weight | 1 |
| Subsampling | 0.8 |
| Column subsampling | 0.7 |

Table 2 – Grid used for hyperparameter search for the first set of models.

The second set of models consisted on XGBoost models trained on the full dataset ($m = 1929$) without hyperparameter tuning (see Table 3 for the hyperparameters used). The models were boosted for $T = 10, 100,$ and 1000 rounds respectively. Each model was then evaluated on the same dataset with respect to predictive value and precision-recall curves. To obtain generalization bands for the predictive values, we applied Theorem 3.3.2 with $\delta = 0.05$ and $\text{VC-sub}(\mathcal{F}_{\text{XGB}}) = 14(T+1)\log_2((T+1)e)$. This approximation is based on the VC dimension of boosting [MRT18] and on the VC dimension of decision stumps (trees of depth 1) [LLM20].

## Empirical experiment

Building on the recommendations outlined in Appendix A.1, we decided to develop the XGBoost model on the basis of a (semi-) nested 5-fold cross-validation scheme

[5] It was also the top performer in [Buz+23].

| Hyperparameter | Values |
|---|---|
| Learning rate | 0.01 |
| Boosting rounds | 10, 100, 1000 |
| Maximum depth | 1 |
| Minimum child weight | 1 |
| Subsampling | 0.8 |
| Column subsampling | 0.7 |

Table 3 – Hyperparameters used for the second set of models.

over the whole dataset, where the optimal number of boosting rounds was determined afresh via 5-fold cross-validation each time the model was fitted to new data (i.e., on a new fold) and the remaining hyperparameters were determined only once at the outer cross-validation scale. To determine the number of boosting rounds, we followed an approach similar to Elith et al. [ELH08], where boosting was performed incrementally until the cross-validation score failed to improve for 200 consecutive rounds[6].

As a consequence of using a semi-nested cross-validation scheme, model selection and validation were performed simultaneously. Specifically, for each choice of hyperparameters considered during model selection, 10 repetitions of 5-fold cross-validation were executed and the mean AUPRC[7] was computed over the resulting 50 folds. The best performing combination of hyperparameters was selected as the final one, and the performance of the model was estimated by recording the precision-recall and calibration curves for each test fold of the winning cross-validation run.

The hyperparameter search was performed using optuna's implementation of the Tree-structured Parzen Estimator (TPE), a bayesian optimization technique that attempts to reduce the number of unnecessary evaluations by keeping track of the "hot regions" of the hyperparameter space [Ber+11]. The hyperparameter search space used is shown in Table 4. To combat overfitting during training, we made sure to include non-trivial values for the subsampling hyperparameters 'subsample' and 'colsample_bytree', to require smaller values of 'max_depth', and to determine 'n_estimators' each time via cross-validation with early stopping. It is important to note that the default values for the other hyperparameters also include an L2 regularization of the tree's weights (see the 'lambda' hyperparameter in the XGBoost documentation[8]).

Although we did not run a fully nested cross-validation as recommended by Cawley and Talbot [CT10] and as such were unable to flag the occurrence of overfitting during model selection, we actively took measure to minimize the chance of it happening. In particular, we reduced the hyperparameter search down to only 3 of the more than 20 hyperparameters available to the tree-based XGBoost model. Moreover, we only ran the

[6] The optimal number of boosting rounds was corrected to account for the difference in training set size from the inner to the outer fold: $T_{\text{final}} = T_{\text{best}}/0.8$.
[7] Computed via scikit-learn's *average_precision_score*.
[8] https://xgboost.readthedocs.io/en/stable/parameter.html

search while the cross-validation score increased at least once every 50 rounds.

| Hyperparameter | Values |
|---|---|
| Learning rate | 0.005 |
| Maximum depth | $\{1, 2, 3, 4\}$ |
| Subsampling | $[0.4, 0.75]$ |
| Column subsampling | $[0.4, 0.75]$ |

Table 4 – Hyperparameter search space used for the optuna study.

### The computational environment

The experiments were performed in Python utilizing the machine learning libraries *scikit-learn* and *xgboost*[9] [CG16; Ped+11]. For data manipulation, we employed *pandas* [McK+11], while visualization was accomplished using the *matplotlib* and *seaborn* libraries [Hun07; Was21]. Additionally, the hyperparameter optimization library *optuna* was used for its implementation of the TPE sampler [Aki+19]. To ensure transparency and reproducibility, we have made the code for all experiments, along with the specific versions and dependencies of the utilized software packages, available in a GitHub Gist[10].

## 4.3 Results

The predictive value and precision-recall curves of the learning-theoretic experiment are shown in Figures 2 and 3. Both figures show the empirical predictive value curves along with the 95% confidence bands derived from Theorems 3.3.1 and 3.3.2.

The optuna study ran for 180 trials and achieved a final mean AUPRC of 0.9527. Figure 4 illustrates the evolution of the model selection criterion over the trials. The final hyperparameters were: 'max_depth' = 2, 'subsample' = 0.6454, and 'colsample_bytree' = 0.7290. The 50 precision-recall and calibration curves of the winning cross-validation run are shown in Figure 5, whereas the histogram of AUPRC values over the 50 folds in shown in Figure 6. The mean precision-recall and calibration curves were computed via linear interpolation with respect to standardized recall and predicted probability values. The confidence bands were determined by considering the standard deviation of the interpolated precision and true probability values. For completeness, Figure 7 shows all four classification curves along with their linearly interpolated means and standard deviations, as discussed in Sections 3.1 and 3.2. Finally, Figure 8 shows the histogram of AUROC (Area Under the Receiver Operator Characteristic curve) values over the 50 folds.

[9] Other similarly good implementations of gradient-boosted trees include *LightGBM* [Ke+17] and *CatBoost* [Pro+18]. Notably, the latter offers native support for categorical features.

[10] https://gist.github.com/alekfrohlich/11a47ce0d19f846e024c0c5602cf60f0.

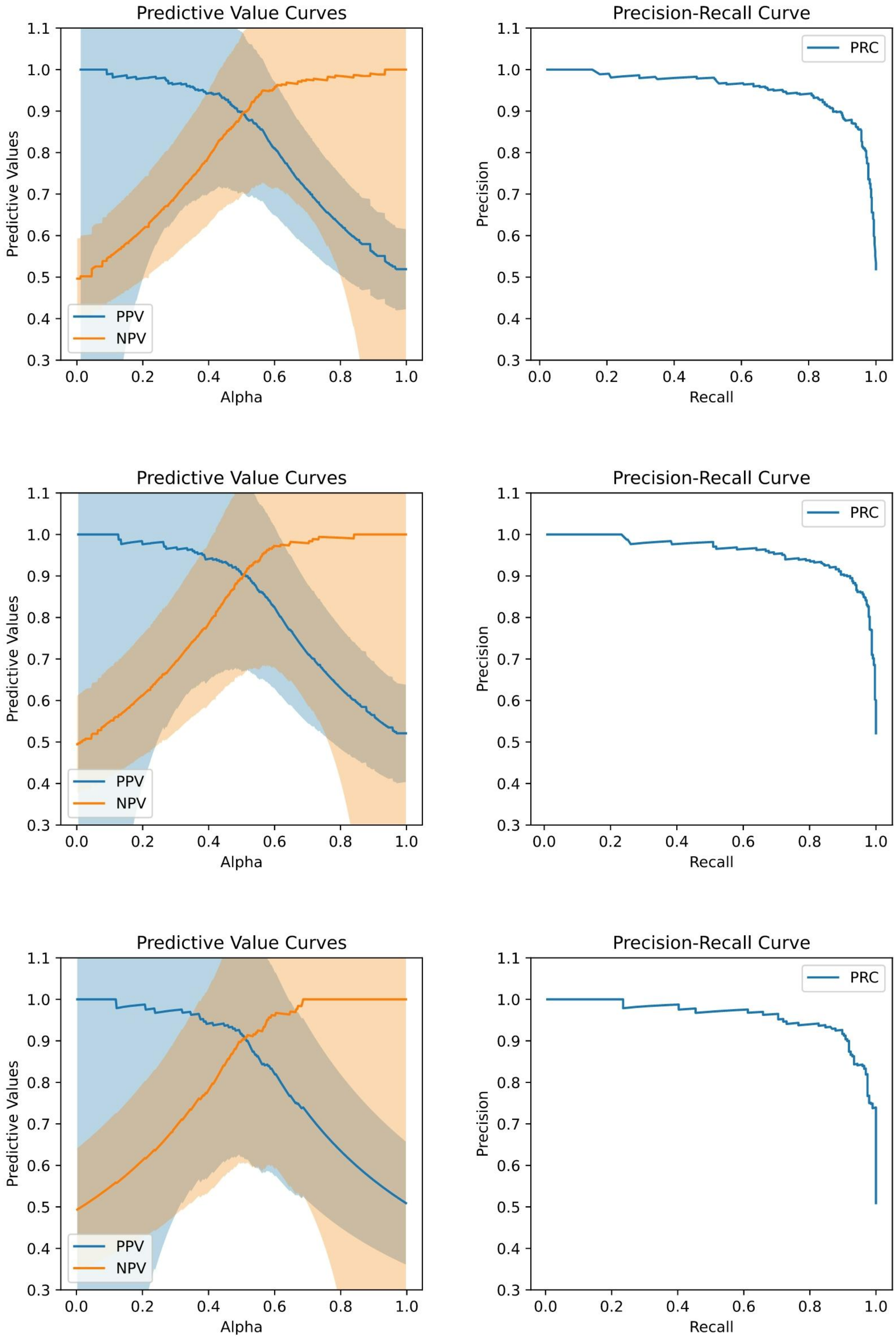


Figure 2 – On the left: empirical predictive value curves along with their 95% confidence bands from Theorem 3.3.1. On the right: empirical precision-recall curves. Each row corresponds to a different train-test split sorted in decreasing order of test set size.

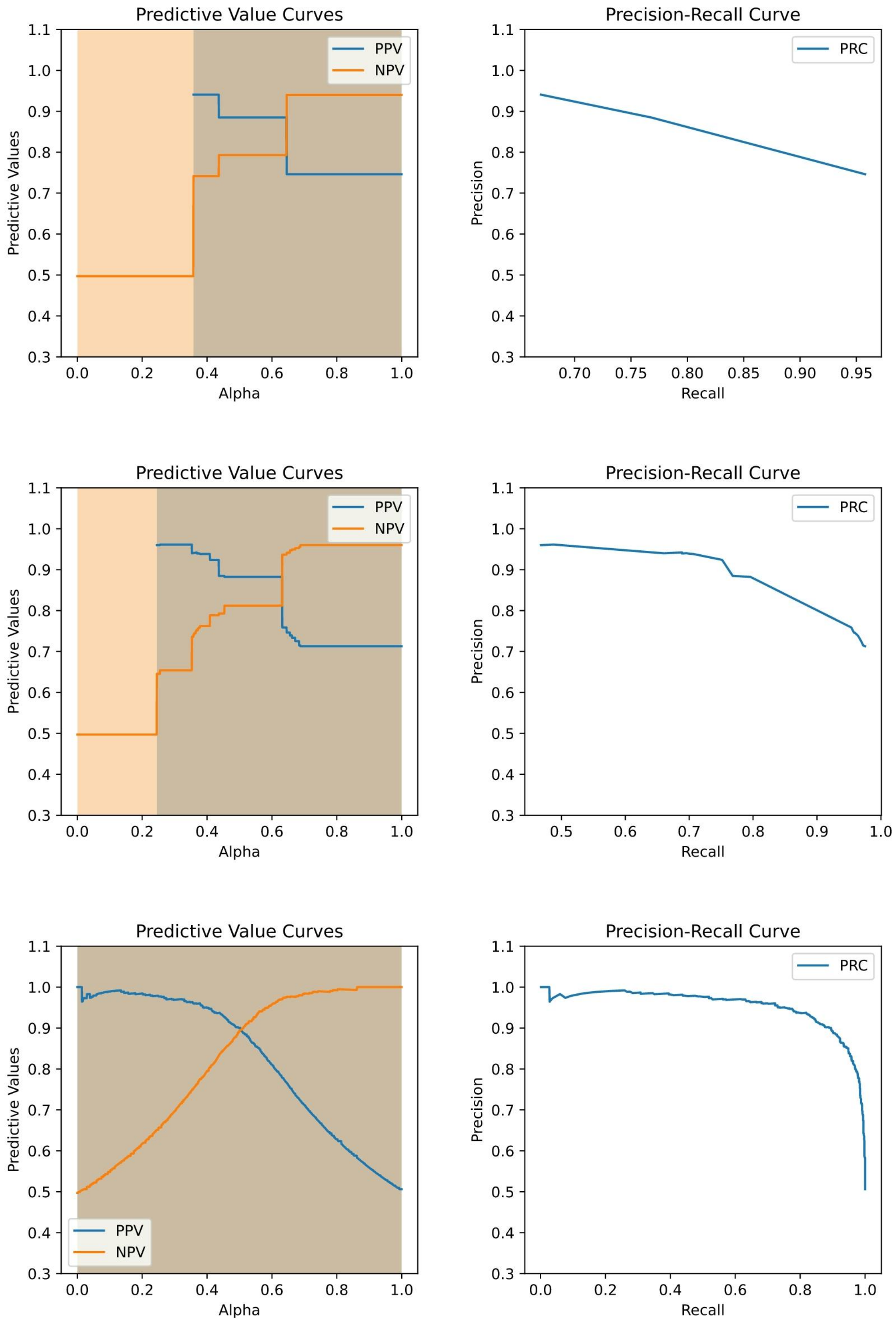


Figure 3 – On the left: empirical predictive value curves along with their 95% confidence bands from Theorem 3.3.2. On the right: empirical precision-recall curves. Each row corresponds to a different number of boosting rounds. From top to bottom: $T = 10$, $T = 100$, and $T = 1000$.

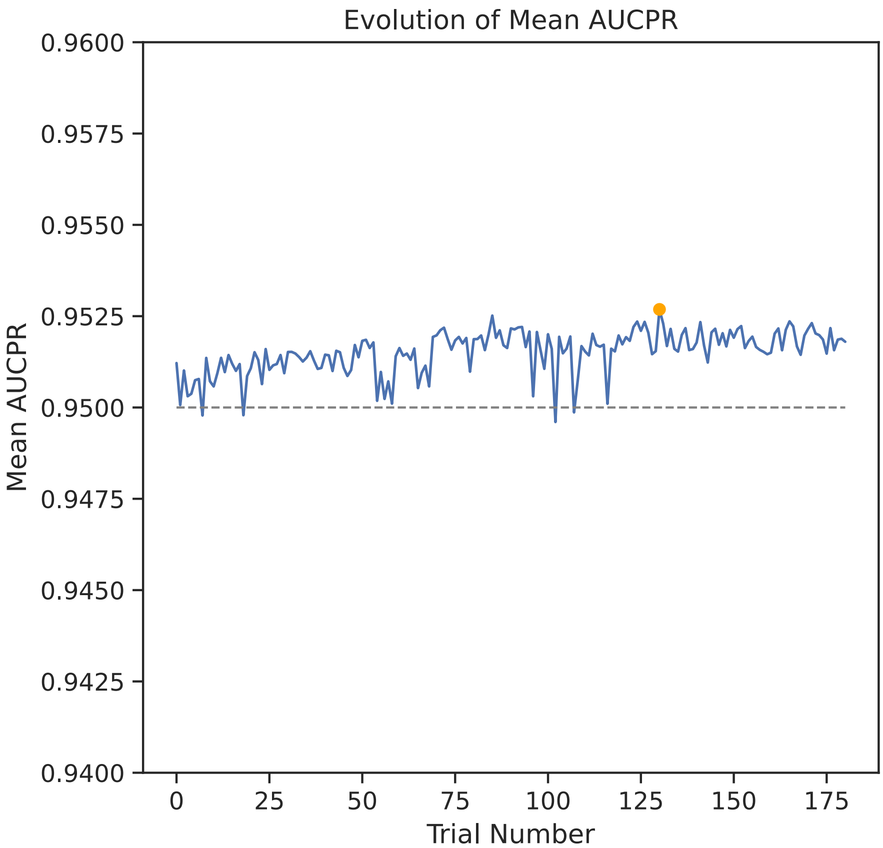


Figure 4 – Evolution of mean AUPRC over the trials. The orange circle represents the best trial.

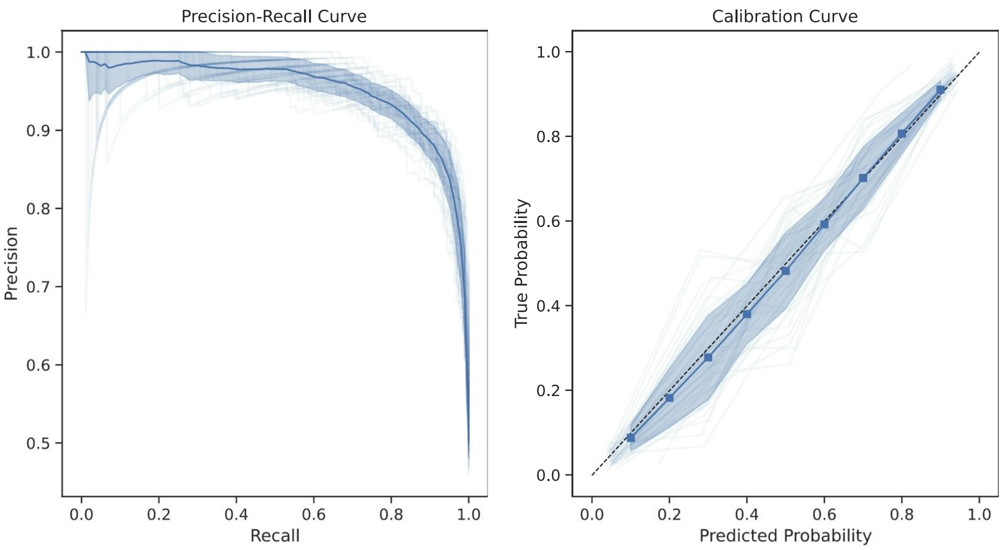


Figure 5 – On the left: precision-recall curves for each test fold, with mean curve and confidence bands computed via linear interpolation. On the right: calibration curves for each test fold, with mean curve and confidence bands computed via linear interpolation.

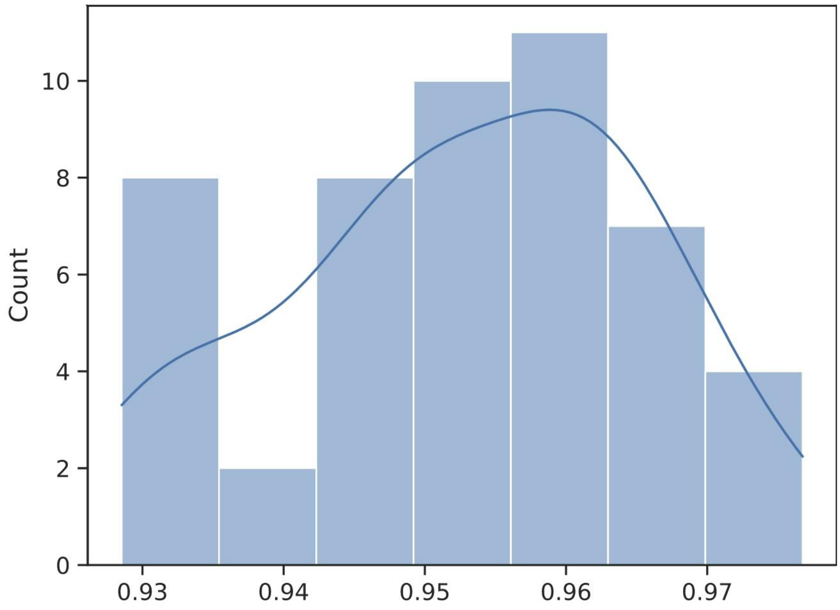


Figure 6 – Histogram of AUPRC values of the winning cross-validation run.

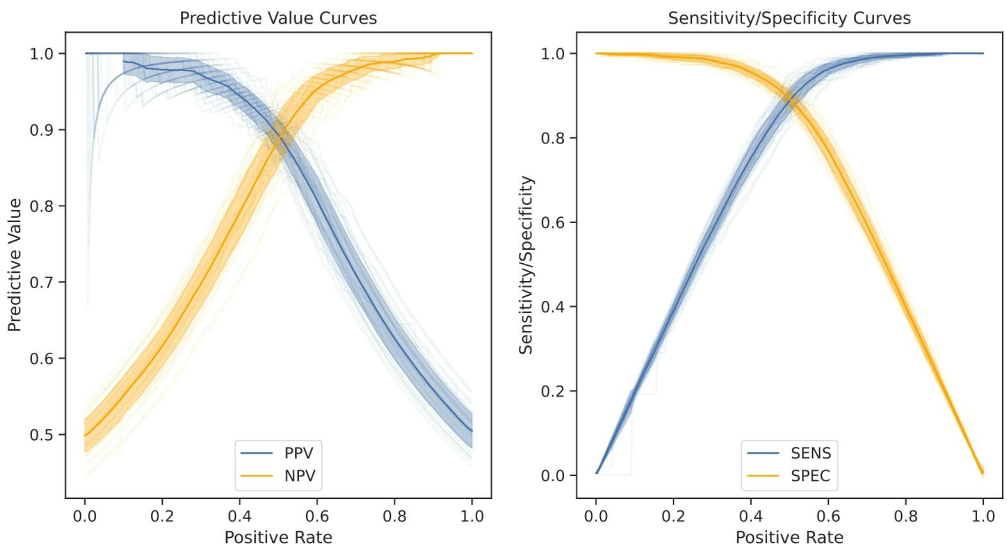


Figure 7 – Predictive value, sensitivity, and specificity curves for each test fold. Mean curves and confidence bands were computed via linear interpolation.

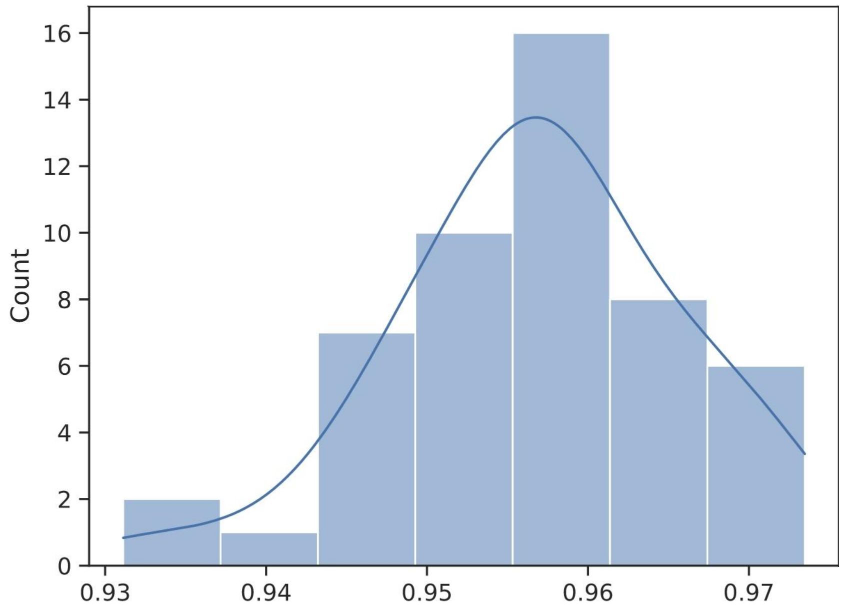


Figure 8 – Histogram of AUROC values of the winning cross-validation run.

### Final model

The final model was obtained by fitting XGBoost with the hyperparameters found in the optuna study over the whole dataset. The final number of trees, determined with 5-fold cross-validation and early stopping, was 1180. The training precision-recall and calibration curves are shown in Figure 9, whereas the training predictive value, sensitivity, and specificity curves are shown in Figure 10. The final model's training AUROC was 0.9623. Individual feature impact over the final model's predictions was computed using TreeExplainer [Lun+20] and is shown in Figure 11 in the form of a density scatter plot of SHAP values, where features are ranked by the sum of absolute SHAP values.

## 4.4 Discussion

It is evident from Figures 2 and 3 that there is simply too much uncertainty present in the confidence bands for them to be useful for model validation. In particular, the uniform bound is completely vacuous due to the high VC dimension of $\mathcal{F}_{\mathrm{XGB}}$. Unfortunately, this situation cannot be ammeliorated as there are no tight estimates for the VC dimension of gradient-boosted trees in the literature, be it the standard VC dimension or the subgraph variant [SF12]. In particular, available estimates for the VC dimension of boosting usually have a linear dependence on the number of boosting rounds $T$, which is commonly set to high values such as $T = 1000$ in practical applications [ELH08]. Moreover, these estimates rely on an upper bound to the VC dimension of the base class, in

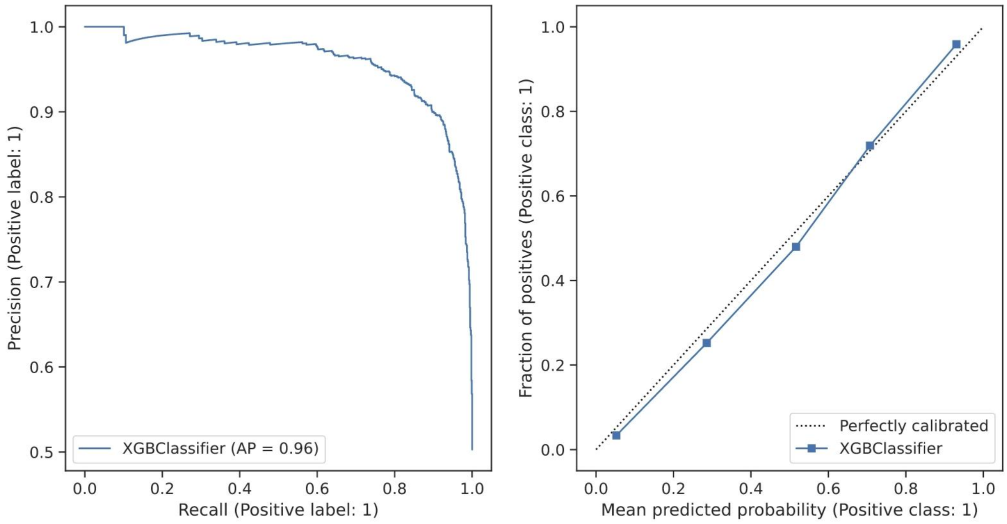


Figure 9 – Final precision-recall and calibration curves computed over the whole dataset.

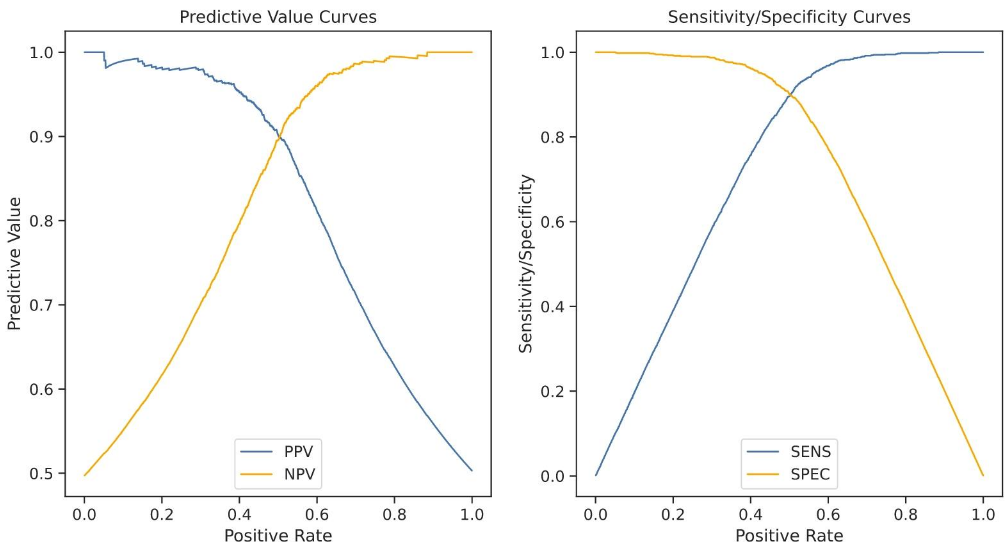


Figure 10 – Final predictive value, sensitivity, and specificity curves computed over the whole dataset.

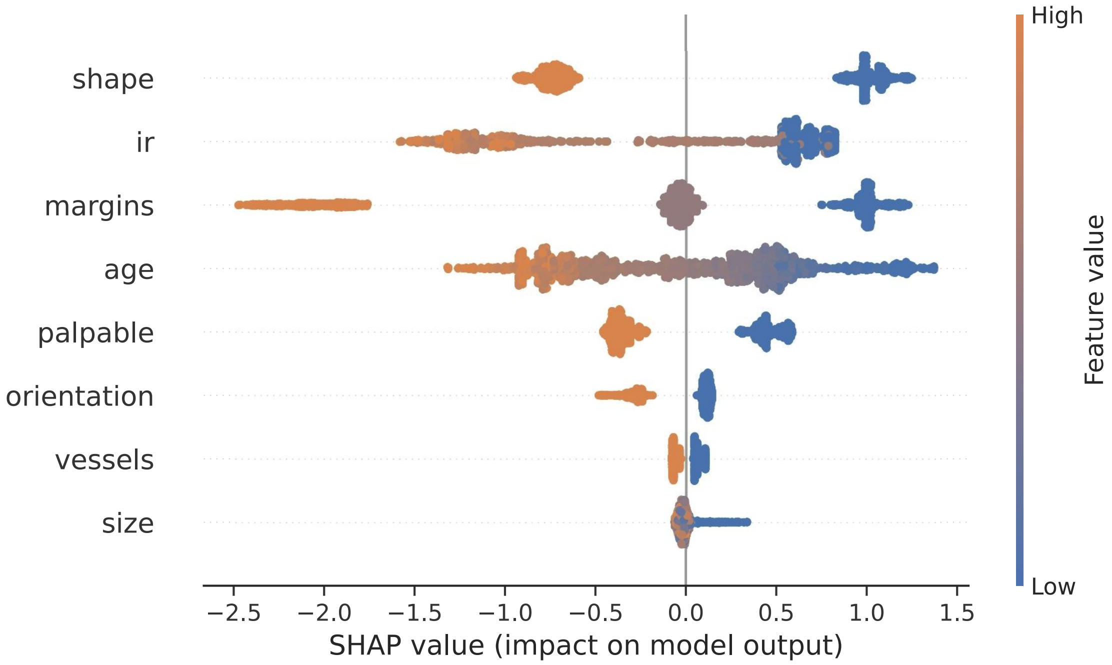


Figure 11 – Density scatter plot of SHAP values computed for the final model over the whole dataset.

our case of VC-dim($\mathcal{H}_{\text{Trees}}(\theta)$). Similar to boosting, the problem of computing the VC dimension of decision trees is still a topic of research, and estimates are typically obtained via recursive procedures based on the VC dimension of low-depth trees [LLM20]. We were able to circumvent this last problem by restricting the base class to trees of depth 1, for which the exact VC dimension is known [LLM20]. Nonetheless, the linear dependence on $T$ still rendered our bounds trivial[11].

Although the bands derived from Theorem 3.3.1 show more promise, they still suffer from a loss of statistical control within the critical range of values $\alpha_k \in (0, 0.5]$. This inherent trade-off between empirical precision and the width of the confidence bands is to be expected, given that a decrease in $\alpha_k$ signifies fewer instances predicted as positive, leading to a diminished sample size from which to compute precision. Regrettably, this scenario forced us to consider validation methods that had a looser connection with the theory introduced in Chapters 2 and 3.

In contrast to the learning-theoretical experiment, the experiment based on cross-validation yielded better results. As can be seen by Figure 4, hyperparameter tuning yielded marginal gains in terms of mean AUPRC. This may be due to the small search space used or due to the fact that the initial hyperparameter values were already close to being optimal. Nonetheless, the final choice of hyperparameters yielded stable precision-recall, ROC, classification, and calibration curves across the 50 folds, as can be seen by Figures 5, 6, 7, and 8. The model achieved good discrimination (mean AUPRC $= 0.9527$) and calibration (mean calibration curve close to $y = x$). The final model similarly achieved

[11] As a matter of fact, for $T = 1000$, our estimate for the VC dimension is already greater than 150000!

good discrimination and calibration on its training set, as Figures 9 and 10 show.

Although we were not able to establish statistical control over the positive predictive value (precision) curve using the predictive value generalization bounds [VS20], we were able to contribute to Buzatto's thesis [Buz24]. In particular, by employing different training and validation procedures to obtain a good model, we strengthen the claim that machine learning may benefit the management of suspicious breast lesions identified by ultrasound. Moreover, the SHAP summary plot shown in Figure 11 corroborates to the importance of multi-modality [Pfo+22], and Doppler features in particular, to the task of malignancy prediction.

In our experiments, we tackled the issue of choosing $\hat{f}_m \in \mathcal{F}$ with good precision-recall, classification, and calibration curves. Contrary to Buzatto et al., we used all the available data for model development and did not commit to any particular trade-off (threshold/positive rate) between the classification metrics. In [Buz+23], a detailed assessment of the final model was performed by evaluating it on a hold-out test set, stratifying by BI-RADS sub-type. The model was able to significantly reduce the number of unnecessary biopsies and had a precision of 98.1%[12]. At this stage, we cannot claim any particular metric values for our model. Nonetheless, it is clear from Figures 5-10 that our model would significantly reduce the number of unnecessary biopsies for multiple threshold levels, including a good part of the critical range $\alpha_k \in (0, 0.5]$ associated with high precision.

[12] The paper doesn't flip the labels, and so by precision we mean negative predictive value.

# 5 Conclusions

In this thesis we studied elements of the mathematical theory of learning. In particular, we reviewed classical concentration inequalities (*Chernoff-Hoeffding* and *McDiarmid*), formalizations of the intuitive notion of model complexity (*VC dimension* and *Rademacher complexity*), and studied the empirical risk minimization principle from an algorithmic perspective. Following this, we formulated the malignancy prediction problem in terms of scoring functions, predictive values (ppv/precision and npv), sensitivity (recall), and specificity, and presented the learning-theoretic results of Vemuri and Srebro [VS20] on the generalization of predictive value curves.

Building upon foundational matters, we proceeded to experiment with the breast lesion dataset collected Buzatto et al. [Buz+23]. In this thesis, we investigated two approaches for developing a gradient boosting model for the prediction of malignancy of breast lesions. The first approach consisted on applying the predictive value generalization bounds to estimate the generalization error of XGBoost models developed on the basis of train-test splits and the whole dataset. Both experiments based on this approach fell short due to insufficient sample sizes and lack of tightness in current estimates for the VC subgraph dimension of gradient-boosted regression trees. The second approach consisted on applying repeated 5-fold cross-validation to select and validate an XGBoost model. This approach worked better and stably yielded models with high discrimination (mean AUPRC = 0.95) and calibration across different cross-validation folds.

There are many possible continuations of this work. First it would be interesting to explore data-dependent generalization bounds for the predictive values and to delve deeper into the issue of estimating the model complexity of boosting in a way that doesn't depend so strongly on the number of boosting rounds, which is typically high for practical applications of boosting. From the perspective of the application, a natural extension of this thesis would be to evaluate the performance of the final model on an external validation set, stratifying by BI-RADS, institution type, age, etc. At this stage, it would also be interesting to analyze individual model predictions more carefully with respect to SHAP values and also within the framework of conformal prediction [AB22], which, although not directly comparable to the learning-theoretic experiments of Chapter 4 [ZYS24], has gained significant traction as a tool for providing distribution-free and model-agnostic uncertainty quantification in high-stakes applications such as healthcare [Csi+23; Lu+22; Sre+24]. Finally, it would be exciting to explore whether a multi-modal (or stacked/ensemble-based) model built using both clinical and image-based features (say, by combining tree ensembles with convolutional neural networks) could outperform our current tabular-based solution.

# Final remarks

We conclude this thesis by highlighting the importance of interdisciplinary research in advancing the current state of healthcare. The field of medicine is full of ripe opportunities for data-driven approaches such as machine learning and inferential statistics. Yet, addressing these challenges requires solutions that incorporate knowledge from multiple domains. In particular, diseases such as cancer display an incredible amount of heterogeneity (viz. [Caj+20]) and so one needs to be very careful with the issues of representativeness and sample size when developing prognostic and diagnostic models for them. Furthermore, one cannot dispense with statistics, and specially mathematical statistics, when considering such intricate problems. This is beatifully put by Bradley Efron [Efr05]:

> A new generation of scientific devices, typified by microarrays, produce data on a gargantuan scale – with millions of data points and thousands of parameters to consider at the same time. These experiments are "deeply statistical". Common sense, and even good scientific intuition, won't do the job by themselves. Careful statistical reasoning is the only way to see through the haze of randomness to the structure underneath. Massive data collection, in astronomy, psychology, biology, medicine, and commerce, is a fact of 21st Century science, and a good reason to buy statistics futures if they are ever offered on the NASDAQ.

At last, one cannot ignore the increased availability of image, text, and video data in the medical field and the increasingly important role played by computer vision, natural language processing, and other deep learning methodologies in dealing with such complex data types (viz. [Pic+21]).

# Contributions

Throughout the development of this thesis, the author participated in the following publications:

1. I. C. Buzatto, S. A. Recife, L. Miguel, N. Onari, A. L. P. Faim, R. M. Bonini, L. Silvestre, D. P. Carlotti, A. Fröhlich, and D. G. Tiezzi. *Machine learning can reliably predict malignancy of breast lesions based on clinical and ultrasonographic features.* 2023. Research Square: 3390199.

2. A. Fröhlich and D. Gonçalves. *SVM: o Problema de Otimização, as Garantias de Aprendizado e o Truque do Kernel.* 2023. Poster presented at the 34th Brazilian Mathematical Colloquium, Rio de Janeiro, Brazil

3. D. Tiezzi, A. Fröhlich, F. Chahud, and S. Pagnota. *197P Computational pathology pipeline enables quantification of intratumor heterogeneity and tumor-infiltrating lymphocyte score.* In: *Immuno-Oncology and Technology* 20 (2023).

4. D. Tiezzi, F. Buono, A. Fröhlich, and S. Pagnotta. *92P Molecular/genomic profile enhances prediction of response to target therapy in HER2-positive breast cancer.* In: *ESMO Open* 9 (2024).

5. A. Fröhlich, A. B. Turcato, and D. G. Tiezzi. *Scientific methodology for descriptive statistics using the R language.* In: *ULakes Journal of Medicine* 3.2 (2023), pages 111–121.

# References


[ABR64] M. Aizerman, E. Braverman, and L. Rozonoer. *Theoretical foundations of the potential function method in pattern recognition learning*. In: *Automation and remote control* 25 (1964), pages 821–837.

[Aki+19] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. *Optuna: A Next-generation Hyperparameter Optimization Framework*. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '19. ACM, 2019.

[AB22] A. N. Angelopoulos and S. Bates. *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. 2022. arXiv: 2107.07511 `[cs.LG]`.

[AC10] S. Arlot and A. Celisse. *A survey of cross-validation procedures for model selection*. In: *Statistics Surveys* 4 (2010), pages 40–79.

[Aro+19] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang. *Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks*. 2019. arXiv: 1901.08584 `[cs.LG]`.

[AL06] K. B. Athreya and S. N. Lahiri. *Measure theory and probability theory*. Springer Texts in Statistics. New York, NY: Springer, 2006.

[AIS18] A. Aydiner, A. Igci, and A. Soran, editors. *Breast cancer*. 1st edition. Cham, Switzerland: Springer International Publishing, 2018.

[BBL02] P. L. Bartlett, S. Boucheron, and G. Lugosi. *Model Selection and Error Estimation*. In: *Machine Learning* 48 (2002), pages 85–113.

[Ben16] Y. Bengio. *Deep Learning*. Adaptive Computation and Machine Learning series. London, England: MIT Press, 2016.

[Ber08] W. A. Berg. *Combined Screening With Ultrasound and Mammography vs Mammography Alone in Women at Elevated Risk of Breast Cancer*. In: *JAMA* 299.18 (2008), pages 2151–2163.

[Ber+11] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. *Algorithms for Hyper-Parameter Optimization*. In: *Advances in Neural Information Processing Systems*. Edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger. Volume 24. Curran Associates, Inc., 2011.

[BGV92] B. E. Boser, I. M. Guyon, and V. N. Vapnik. *A training algorithm for optimal margin classifiers*. In: *Proceedings of the fifth annual workshop on Computational learning theory*. COLT92. ACM, 1992.

[BLM16] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities. A Nonasymptotic Theory of Independence*. London, England: Oxford University Press, 2016.

[Bra+24] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal. *Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. In: *CA: A Cancer Journal for Clinicians* (2024).

[Bre01] L. Breiman. *Random Forests*. In: *Machine Learning* 45.1 (2001), pages 5–32.

[Buz24] I. C. Buzatto. *Machine learning como metodo de predição de malignidade em lesões mamarias identificadas à ultrassonografia*. PhD Thesis. Faculdade de Medicina de Ribeirão Preto, USP, 2024.

[Buz+23] I. C. Buzatto, S. A. Recife, L. Miguel, N. Onari, A. L. P. Faim, R. M. Bonini, L. Silvestre, D. P. Carlotti, A. Fröhlich, and D. G. Tiezzi. *Machine learning can reliably predict malignancy of breast lesions based on clinical and ultrasonographic features*. 2023. Research Square: 3390199.

[Caj+20] S. R. y Cajal, M. Sesé, C. Capdevila, T. Aasen, L. D. Mattos-Arruda, S. J. Diaz-Cano, J. Hernández-Losa, and J. Castellví. *Clinical implications of intratumor heterogeneity: challenges and opportunities*. In: *Journal of Molecular Medicine* 98.2 (2020), pages 161–177.

[Car23] H. Carlens. *State of Competitive Machine Learning in 2022*. In: *ML Contests Research* (2023).

[CT10] G. C. Cawley and N. L. C. Talbot. *On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation*. In: *Journal of Machine Learning Research* 11.70 (2010), pages 2079–2107.

[CG16] T. Chen and C. Guestrin. *XGBoost: A Scalable Tree Boosting System*. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD'16. ACM, 2016.

[Che52] H. Chernoff. *A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations*. In: *The Annals of Mathematical Statistics* 23.4 (1952), pages 493–507.

[Cho+19] J. S. Choi, B.-K. Han, E. S. Ko, J. M. Bae, E. Y. Ko, S. H. Song, M.-r. Kwon, J. H. Shin, and S. Y. Hahn. *Effect of a Deep Learning Framework-Based Computer-Aided Diagnosis System on the Diagnostic Performance of Radiologists in Differentiating between Malignant and Benign Masses on Breast Ultrasonography*. In: *Korean Journal of Radiology* 20.5 (2019), pages 749–758.

[Cor+11] V. Corsetti, N. Houssami, M. Ghirardi, A. Ferrari, M. Speziani, S. Bellarosa, G. Remida, C. Gasparotti, E. Galligioni, and S. Ciatto. *Evidence of the effect of adjunct ultrasound screening in women with mammography-negative dense breasts: Interval breast cancers at 1year follow-up*. In: *European Journal of Cancer* 47.7 (2011), pages 1021–1026.

[CV95] C. Cortes and V. Vapnik. *Support-vector networks*. In: *Machine Learning* 20.3 (1995), pages 273–297.

[CT18] H. Cramer and H. Touchette. *On a new limit theorem in probability theory (Translation of 'Sur un nouveau theoreme-limite de la theorie des probabilites')*. 2018. arXiv: 1802.05988 [math.HO].

[Cra03] J. Cramer. *The Origins of Logistic Regression*. In: *SSRN Electronic Journal* (2003).

[Csi+23] D. Csillag, L. Monteiro Paes, T. Ramos, J. V. Romano, R. Schuller, R. B. Seixas, R. I. Oliveira, and P. Orenstein. *AmnioML: Amniotic Fluid Segmentation and Volume Prediction with Uncertainty Quantification*. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.13 (2023), pages 15494–15502.

[Dar+15] T. Darrell, M. Kloft, M. Pontil, G. Rätsch, and E. Rodner. *Machine Learning with Interdependent and Non-identically Distributed Data*. Report from Dagstuhl Seminar 15152. Dagstuhl, 2015.

[DFO20] M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for machine learning*. Cambridge, England: Cambridge University Press, 2020.

[Deo15] R. C. Deo. *Machine Learning in Medicine*. In: *Circulation* 132.20 (2015), pages 1920–1930.

[Doo40] J. L. Doob. *Regularity properties of certain families of chance variables*. In: *Transactions of the American Mathematical Society* 47.3 (1940), pages 455–486.

[Dos+20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2020. arXiv: 2010.11929 [cs.CV].

[DR17] G. K. Dziugaite and D. M. Roy. *Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data*. In: (2017). arXiv: 1703.11008 [cs.LG].

[Efr79] B. Efron. *Bootstrap Methods: Another Look at the Jackknife*. In: *The Annals of Statistics* 7.1 (1979), pages 1–26.

[Efr83] B. Efron. *Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation*. In: *Journal of the American Statistical Association* 78.382 (1983), pages 316–331.

[Efr05] B. Efron. *Modern science and the Bayesian-frequentist controversy*. Division of Biostatistics, Stanford University Stanford, CA, USA, 2005.

[ELH08] J. Elith, J. R. Leathwick, and T. Hastie. *A working guide to boosted regression trees*. In: *Journal of Animal Ecology* 77.4 (2008), pages 802–813.

[Eva+18] A. Evans, R. M. Trimboli, A. Athanasiou, C. Balleyguier, P. A. Baltzer, U. Bick, J. Camps Herrero, P. Clauser, C. Colin, E. Cornford, E. M. Fallenberg, M. H. Fuchsjaeger, F. J. Gilbert, T. H. Helbich, K. Kinkel, S. H. Heywang-Köbrunner, C. K. Kuhl, R. M. Mann, L. Martincich, P. Panizza, F. Pediconi, R. M. Pijnappel, K. Pinker, S. Zackrisson, G. Forrai, and F. Sardanelli. *Breast ultrasound: recommendations for information to women and referring physicians by the European Society of Breast Imaging*. In: *Insights into Imaging* 9.4 (2018), pages 449–461.

[Fei10] S. Feig. *Cost-Effectiveness of Mammography, MRI, and Ultrasonography for Breast Cancer Screening*. In: *Radiologic Clinics of North America* 48.5 (2010), pages 879–891.

[FS97] Y. Freund and R. E. Schapire. *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*. In: *Journal of Computer and System Sciences* 55.1 (1997), pages 119–139.

[FHT00] J. Friedman, T. Hastie, and R. Tibshirani. *Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)*. In: *The Annals of Statistics* 28.2 (2000), pages 337–407.

[Fri01] J. H. Friedman. *Greedy function approximation: A gradient boosting machine*. In: *The Annals of Statistics* 29.5 (2001), pages 1189–1232.

[FG23] A. Fröhlich and D. Gonçalves. *SVM: o Problema de Otimização, as Garantias de Aprendizado e o Truque do Kernel*. 2023. Poster presented at the 34th Brazilian Mathematical Colloquium, Rio de Janeiro, Brazil.

[FTT23] A. Fröhlich, A. B. Turcato, and D. G. Tiezzi. *Scientific methodology for descriptive statistics using the R language*. In: *ULakes Journal of Medicine* 3.2 (2023), pages 111–121.

[Ger+24] C. Gerbelot, A. Karagulyan, S. Karp, K. Ravichandran, M. Stern, and N. Srebro. *Applying statistical learning theory to deep learning*. 2024. arXiv: 2311.15404 `[cs.LG]`.

[GOV22] L. Grinsztajn, E. Oyallon, and G. Varoquaux. *Why do tree-based models still outperform deep learning on typical tabular data?* In: *Advances in Neural Information Processing Systems*. Edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Volume 35. Curran Associates, Inc., 2022, pages 507–520.

[GE03] I. Guyon and A. Elisseeff. *An introduction to variable and feature selection*. In: *Journal of machine learning research* 3 (2003), pages 1157–1182.

[Har15] F. E. Harrell. *Regression modeling strategies*. 2nd edition. Springer Series in Statistics. Cham, Switzerland: Springer International Publishing, 2015.

[Har19] F. E. Harrell. *Classification vs. Prediction*. Last accessed on March 2024. 2019. URL: `https://www.fharrell.com/post/classification/`.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York: Springer New York, 2009.

[Hoe63] W. Hoeffding. *Probability Inequalities for Sums of Bounded Random Variables*. In: *Journal of the American Statistical Association* 58.301 (1963), pages 13–30.

[Hon+22] A. A. H. de Hond, A. M. Leeuwenberg, L. Hooft, I. M. J. Kant, S. W. J. Nijman, H. J. A. van Os, J. J. Aardoom, T. P. A. Debray, E. Schuit, M. van Smeden, J. B. Reitsma, E. W. Steyerberg, N. H. Chavannes, and K. G. M. Moons. *Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review*. In: *npj Digital Medicine* 5.2 (2022).

[Hun07] J. D. Hunter. *Matplotlib: A 2D graphics environment*. In: *Computing in Science & Engineering* 9.3 (2007), pages 90–95.

[IMS02] IMS. *The Publications and Writings of John W. Tukey*. In: *The Annals of Statistics* 30.6 (2002), pages 1666–1680.

[INC23] INCA. *Dados e Numeros sobre Câncer de Mama - Relatório Anual 2023*. Instituto Nacional de Câncer, 2023.

[Jac+21] R. Jacobucci, A. K. Littlefield, A. J. Millner, E. M. Kleiman, and D. Steinley. *Evidence of Inflated Prediction Performance: A Commentary on Machine Learning and Suicide Research*. In: *Clinical Psychological Science* 9.1 (2021), pages 129–134.

[Ke+17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. In: *Advances in Neural Information Processing Systems*. Edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Volume 30. Curran Associates, Inc., 2017.

[KV94] M. Kearns and L. Valiant. *Cryptographic limitations on learning Boolean formulae and finite automata*. In: *Journal of the ACM* 41.1 (1994), pages 67–95.

[Kim+21] S.-Y. Kim, Y. Choi, E.-K. Kim, B.-K. Han, J. H. Yoon, J. S. Choi, and J. M. Chang. *Deep learning-based computer-aided diagnosis in screening breast ultrasound to reduce false-positive diagnoses*. In: *Scientific Reports* 11.395 (2021).

[Koh95] R. Kohavi. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995, pages 1137–1143.

[Kol01] V. Koltchinskii. *Rademacher penalties and structural risk minimization*. In: *IEEE Transactions on Information Theory* 47.5 (2001), pages 1902–1914.

[KP00] V. Koltchinskii and D. Panchenko. *Rademacher Processes and Bounding the Risk of Function Learning*. In: *High Dimensional Probability II*. Birkhäuser Boston, 2000, pages 443–457.

[Kos+22] V. Kostic, P. Novelli, A. Maurer, C. Ciliberto, L. Rosasco, and M. Pontil. *Learning Dynamical Systems via Koopman Operator Regression in Reproducing Kernel Hilbert Spaces*. 2022. arXiv: 2205.14027 `[cs.LG]`.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. In: *Advances in Neural Information Processing Systems*. Edited by F. Pereira, C. Burges, L. Bottou, and K. Weinberger. Volume 25. Curran Associates, Inc., 2012.

[Kum+22] V. Kumar, A. K. Abbas, J. C. Aster, and A. T. Deyrup, editors. *Robbins & Kumar basic pathology*. 11th edition. Robbins Pathology. Philadelphia, PA: Elsevier - Health Sciences Division, 2022.

[LM68] P. A. Lachenbruch and M. R. Mickey. *Estimation of Error Rates in Discriminant Analysis*. In: *Technometrics* 10 (1968), pages 1–11.

[LLM20] J.-S. Leboeuf, F. LeBlanc, and M. Marchand. *Decision trees as partitioning machines to characterize their generalization properties*. 2020. arXiv: 2010.07374 `[cs.LG]`.

[LeC+89] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. *Backpropagation Applied to Handwritten Zip Code Recognition*. In: *Neural Computation* 1.4 (1989), pages 541–551.

[Lee+12] J.-H. Lee, Y. K. Seong, C.-H. Chang, J. Park, M. Park, K.-G. Woo, and E. Y. Ko. *Fourier-based shape feature extraction technique for computer-aided B-Mode ultrasound diagnosis of breast tumor*. In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012.

[Lu+22] C. Lu, K. Chang, P. Singh, and J. Kalpathy-Cramer. *Three Applications of Conformal Prediction for Rating Breast Density in Mammography*. 2022. arXiv: 2206.12008 `[eess.IV]`.

[Lun+20] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. *From local explanations to global understanding with explainable AI for trees*. In: *Nature Machine Intelligence* 2.1 (2020), pages 56–67.

[Mas00] P. Massart. *Some applications of concentration inequalities to statistics*. In: *Annales de la faculte des sciences de Toulouse Mathematiques* 9.2 (2000), pages 245–303.

[MP43] W. S. McCulloch and W. Pitts. *A logical calculus of the ideas immanent in nervous activity*. In: *The Bulletin of Mathematical Biophysics* 5.4 (1943), pages 115–133.

[McK+11] W. McKinney et al. *pandas: a foundational Python library for data analysis and statistics*. In: *Python for high performance and scientific computing* 14.9 (2011), pages 1–9.

[MBB+13] E. Mendelson, M. Böhm-Vélez, W. Berg, et al. *ACR BI-RADS Ultrasound*. In: ACR BI-RADS Atlas, Breast Imaging Reporting and Data System: American College of Radiology, 2013.

[MRT18] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. 2nd edition. Adaptive Computation and Machine Learning series. London, England: MIT Press, 2018.

[MT68] F. Mosteller and J. Tukey. *Data analysis, including statistics*. In: *Handbook of Social Psychology* (1968).

[Mou+20] A. F. Moustafa, T. W. Cary, L. R. Sultan, S. M. Schultz, E. F. Conant, S. S. Venkatesh, and C. M. Sehgal. *Color Doppler Ultrasound Improves Machine Learning Diagnosis of Breast Cancer*. In: *Diagnostics* 10.9 (2020), page 631.

[Nie16] D. Nielsen. *Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition?* Master's Thesis. Norwegian University of Science and Technology, 2016.

[Nov62] A. B. Novikoff. *On convergence proofs on perceptrons.* In: *Proceedings of the Symposium on the Mathematical Theory of Automata.* Volume 12. 1. New York, NY. 1962, pages 615–622.

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. *Scikit-learn: Machine learning in Python.* In: *the Journal of Machine Learning Research* 12 (2011), pages 2825–2830.

[Pfo+22] A. Pfob, C. Sidey-Gibbons, R. G. Barr, V. Duda, Z. Alwafai, C. Balleyguier, D.-A. Clevert, S. Fastner, C. Gomez, M. Goncalo, I. Gruber, M. Hahn, A. Hennigs, P. Kapetas, S.-C. Lu, J. Nees, R. Ohlinger, F. Riedel, M. Rutten, B. Schaefgen, M. Schuessler, A. Stieber, R. Togawa, M. Tozaki, S. Wojcinski, C. Xu, G. Rauch, J. Heil, and M. Golatta. *The importance of multi-modal imaging and clinical information for humans and AI-based algorithms to classify breast masses (INSPiRED 003): an international, multicenter analysis.* In: *European Radiology* 32.6 (2022), pages 4101–4115.

[Pic+21] F. Piccialli, V. D. Somma, F. Giampaolo, S. Cuomo, and G. Fortino. *A survey on deep learning in medicine: Why, how and when?* In: *Information Fusion* 66 (2021), pages 111–137.

[Pla+99] J. Platt et al. *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.* In: *Advances in large margin classifiers* 10.3 (1999), pages 61–74.

[Pra+18] G. W. Prager, S. Braga, B. Bystricky, C. Qvortrup, C. Criscitiello, E. Esin, G. S. Sonke, G. Martinez, J.-S. Frenel, M. Karamouzis, M. Strijbos, O. Yazici, P. Bossi, S. Banerjee, T. Troiani, A. Eniu, F. Ciardiello, J. Tabernero, C. C. Zielinski, P. G. Casali, F. Cardoso, J.-Y. Douillard, S. Jezdic, K. McGregor, G. Bricalli, M. Vyas, and A. Ilbawi. *Global cancer control: responding to the growing burden, rising costs and inequalities in access.* In: *ESMO Open* 3.2 (2018), e000285.

[Pro+18] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. *CatBoost: unbiased boosting with categorical features.* In: *Advances in Neural Information Processing Systems.* Edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Volume 31. Curran Associates, Inc., 2018.

[Rad21] J. Radon. *Mengen konvexer Körper, die einen gemeinsamen Punkt enthalten.* In: *Mathematische Annalen* 83.1–2 (1921), pages 113–115.

[RDK19] A. Rajkomar, J. Dean, and I. Kohane. *Machine learning in medicine.* In: *New England Journal of Medicine* 380.14 (2019), pages 1347–1358.

[Ros62] F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Washington, D.C.: Spartan Books, 1962.

[Sau72] N. Sauer. *On the density of families of sets.* In: *Journal of Combinatorial Theory, Series A* 13.1 (1972), pages 145–147.

[SF12] R. E. Schapire and Y. Freund. *Boosting: Foundations and Algorithms.* Adaptive Computation And Machine Learning Series. London, England: MIT Press, 2012.

[SS18] B. Schoelkopf and A. J. Smola. *Learning with kernels.* Adaptive Computation and Machine Learning series. London, England: MIT Press, 2018.

[She+07] W.-C. Shen, R.-F. Chang, W. K. Moon, Y.-H. Chou, and C.-S. Huang. *Breast Ultrasound Computer-Aided Diagnosis Using BI-RADS Features.* In: *Academic Radiology* 14.8 (2007), pages 928–939.

[She+23] Y. Shen, J. Park, F. Yeung, E. Goldberg, L. Heacock, F. Shamout, and K. J. Geras. *Leveraging Transformers to Improve Breast Cancer Classification and Risk Assessment with Multi-modal and Longitudinal Data.* 2023. arXiv: 2311.03217 `[eess.IV]`.

[She+21] Y. Shen, F. E. Shamout, J. R. Oliver, J. Witowski, K. Kannan, J. Park, N. Wu, C. Huddleston, S. Wolfson, A. Millet, R. Ehrenpreis, D. Awal, C. Tyma, N. Samreen, Y. Gao, C. Chhor, S. Gandhi, C. Lee, S. Kumari-Subaiya, C. Leonard, R. Mohammed, C. Moczulski, J. Altabet, J. Babb, A. Lewin, B. Reig, L. Moy, L. Heacock, and K. J. Geras. *Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams.* In: *Nature Communications* 12.1 (2021).

[Sie+23] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal. *Cancer statistics, 2023.* In: *CA: A Cancer Journal for Clinicians* 73.1 (2023), pages 17–48.

[SBC20] R. T. Sivarajah, K. Brown, and A. Chetlen. *"I can see clearly now." fundamentals of breast ultrasound optimization.* In: *Clinical Imaging* 64 (2020), pages 124–135.

[SNW11] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning.* Neural Information Processing series. London, England: MIT Press, 2011.

[Sre+24] A. P. Sreenivasan, A. Vaivade, Y. Noui, P. E. Khoonsari, J. Burman, O. Spjuth, and K. Kultima. *Conformal prediction enables disease course prediction and allows individualized diagnostic uncertainty in multiple sclerosis.* In: *medRxiv* (2024). eprint: `https://www.medrxiv.org/content/early/2024/03/02/2024.03.01.24303566.full.pdf`.

[Ste+01] E. W. Steyerberg, F. E. Harrell, G. J. Borsboom, M. Eijkemans, Y. Vergouwe, and J. F. Habbema. *Internal validation of predictive models*. In: *Journal of Clinical Epidemiology* 54.8 (2001), pages 774–781.

[Ste18] E. W. Steyerberg. *Validation in prediction research: the waste by data splitting*. In: *Journal of Clinical Epidemiology* 103 (2018), pages 131–133.

[Ste19] E. W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health. Cham, Switzerland: Springer Cham, 2019.

[Sto74] M. Stone. *Cross-Validatory Choice and Assessment of Statistical Predictions*. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pages 111–133.

[SWM93] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. *Nuclear feature extraction for breast tumor diagnosis*. In: *Biomedical Image Processing and Biomedical Visualization*. Edited by R. S. Acharya and D. B. Goldgof. SPIE, 1993.

[Tha20] A. Tharwat. *Classification assessment methods*. In: *Applied computing and informatics* 17.1 (2020), pages 168–192.

[Tie+24] D. Tiezzi, F. Buono, A. Fröhlich, and S. Pagnotta. *92P Molecular/genomic profile enhances prediction of response to target therapy in HER2-positive breast cancer*. In: *ESMO Open* 9 (2024).

[Tie+23a] D. Tiezzi, A. Fröhlich, F. Chahud, and S. Pagnota. *197P Computational pathology pipeline enables quantification of intratumor heterogeneity and tumor-infiltrating lymphocyte score*. In: *Immuno-Oncology and Technology* 20 (2023).

[Tie+23b] D. Tiezzi, A. Frohlich, F. Chahud, and S. Pagnotta. *197P Computational pathology pipeline enables quantification of intratumor heterogeneity and tumor-infiltrating lymphocyte score*. In: *Immuno-Oncology and Technology* (2023).

[Tre17] R. Trevethan. *Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice*. In: *Frontiers in Public Health* 5.307 (2017).

[VW23] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. 2nd edition. Cham, Switzerland: Springer International Publishing, 2023.

[Val84] L. G. Valiant. *A theory of the learnable*. In: *Communications of the ACM* 27.11 (1984), pages 1134–1142.

[VC68] V. N. Vapnik and A. Y. Chervonenkis. *Uniform convergence of frequencies of occurence of events to their probabilities*. English. In: *Sov. Math., Dokl.* 9 (1968), pages 915–918.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley-interscience, 1998.

[Vap10] V. Vapnik. *The nature of statistical learning theory*. Information Science and Statistics. New York, NY: Springer, 2010.

[VC74] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Moscow: Nauka, 1974.

[VC89] V. Vapnik and A. Chervonenkis. *The necessary and sufficient conditions for consistency of the method of empirical risk minimization*. In: *Pattern Recognition and Image Analysis* 1.3 (1989), pages 283–305.

[VS20] K. Vemuri and N. Srebro. *Predictive Value Generalization Bounds*. 2020. arXiv: 2007.05073 `[stat.ML]`.

[Was21] M. L. Waskom. *Seaborn: statistical data visualization*. In: *Journal of Open Source Software* 6.60 (2021), page 3021.

[Yan+20] L. Yang, S. Wang, L. Zhang, C. Sheng, F. Song, P. Wang, and Y. Huang. *Performance of ultrasonography screening for breast cancer: a systematic review and meta-analysis*. In: *BMC Cancer* 20.1 (2020).

[Zel+19] J. C. van Zelst, T. Tan, R. M. Mann, and N. Karssemeijer. *Validation of radiologists' findings by computer-aided detection (CAD) software in breast cancer detection with automated 3D breast ultrasound: a concept study in implementation of artificial intelligence software*. In: *Acta Radiologica* 61.3 (2019), pages 312–320.

[Zha+16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. *Understanding deep learning requires rethinking generalization*. 2016. arXiv: 1611.03530 `[cs.LG]`.

[ZYS24] Z. Zheng, B. Yang, and P. Song. *Quantifying Uncertainty in Classification Performance: ROC Confidence Bands Using Conformal Prediction*. 2024. arXiv: 2405.12953 `[stat.ME]`.

# A Appendix

## A.1 Empirical studies on model selection and validation

Perhaps the simplest strategy for model selection and validation would be to train multiple models on the full dataset and then select and report the metrics of the model with the best training performance. However, such an approach is known to lead to overoptimistic results [Ste+01]. Recommended solutions involve either penalizing the training metric by the complexity of the solution found [MRT18, Chapter 4] or by its estimated optimism [HTF09, Chapter 7]. As is shown in Section 4.3, the first approach is not compatible with our chosen methodology, given that current estimates for the VC dimension of gradient-boosted trees have too strong of a dependence on the number of boosting rounds and our dataset size is too small for the available generalization bounds to be useful.

A commonly offered alternative solution from the statistical community is to estimate the optimism through bootstrapping, resulting in a procedure known as the optimism-corrected bootstrap [HTF09; Ste+01]. The procedure consists in resampling the dataset with replacement, then training a model on this bootstrapped sample, and finally recording the performance of the bootstrapped model on both the bootstrapped sample and the full dataset. The difference in performance between the two samples is considered an estimate of the optimism. This procedure is usually repeated multiple times and then averaged for increased stability.

Although a powerful tool for simple models such as linear regression, optimism-corrected bootstrap may be ineffective when combined with modern machine learning methods such neural networks[1], ensembles of trees (see Figure 12), and support vector machines. For instance, Jacobucci et al. [Jac+21] shows that such technique may be specially dangerous when combined with random forests or gradient boosting and puts into question the reliability of the results obtained using this strategy in the context of suicide research in psychology.

A second intuitive strategy is the train-test split. It consists in randomly partitioning the dataset into two subsets: a training and a testing set. In this scenario, model selection could be performed by training multiple models on the training set and then choosing the best performing model on the test set. Model validation would then be performed by recording the metrics of the chosen model on the test set.

This approach may also lead to overoptimistic results, as one is implicitly re-

[1] For an interesting example of a convolutional neural network being able to interpolate a dataset of roughly one million images distributed across a thousand distinct classes, see [Zha+16].

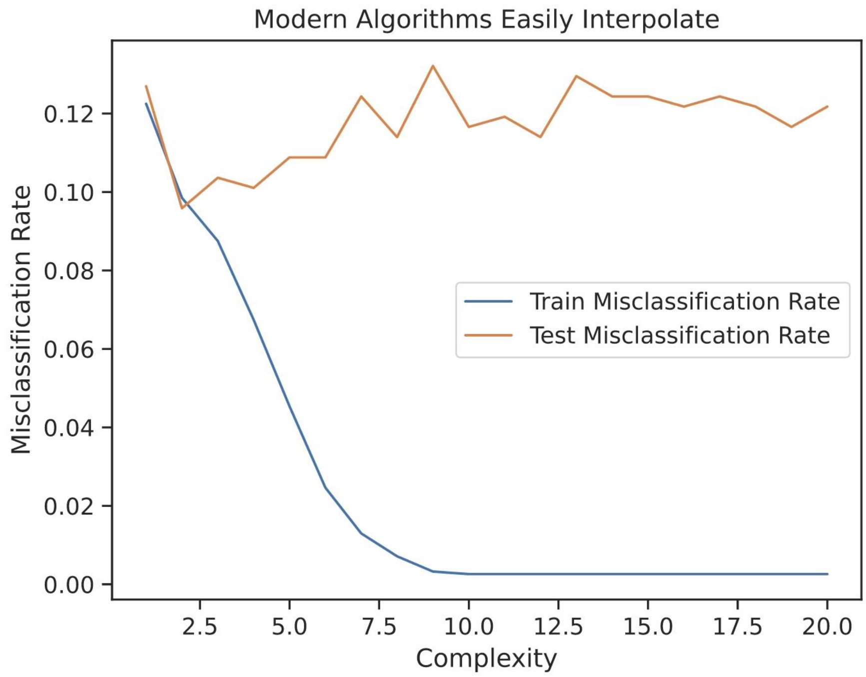


Figure 12 – Behavior of the train and test error curves of different XGBoost models trained on 80% of the breast lesion dataset ($m = 1543$) as a function of their complexity. The models were each trained using the default hyperparameters, except for 'n_estimators' and 'max_depth' which were respectively set to $10x$ and $x$ for $x \in [20]$.

lying on random fluctuations in the performance of different models as evidence for increased generalizability. See the following quote from *The Elements of Statistical Learning* [HTF09, p. 222]

> Ideally, the test set should be kept in a "vault", and be brought out only at the end of the data analysis. Suppose instead that we use the test-set repeatedly, choosing the model with smallest test-set error. Then the test set error of the final chosen model will underestimate the true test error, sometimes substantially.

This issue may be corrected by either performing a triple split into training, validation, and test sets, or by performing model selection on the training set. For the second approach, one might use either the penalized training performance or the mean cross-validated (or bootstrapped) score as a model selection criterion. Nonetheless, the usage of the split is frowned upon by some statisticians working in the development of clinical prediction models. According to Steyerberg [Ste18], the procedure is at best wasteful, given that data splitting only uses part of the available data for model development. Moreover, he argues that the procedure might lead to weak validation studies in scenarios involving very small datasets. Harrell [Har15] maintains that data splitting is wasteful, but adds that the procedure masks the inherent variability present in modeling with small datasets, which are very common in biomedical research (see Figure 13). In

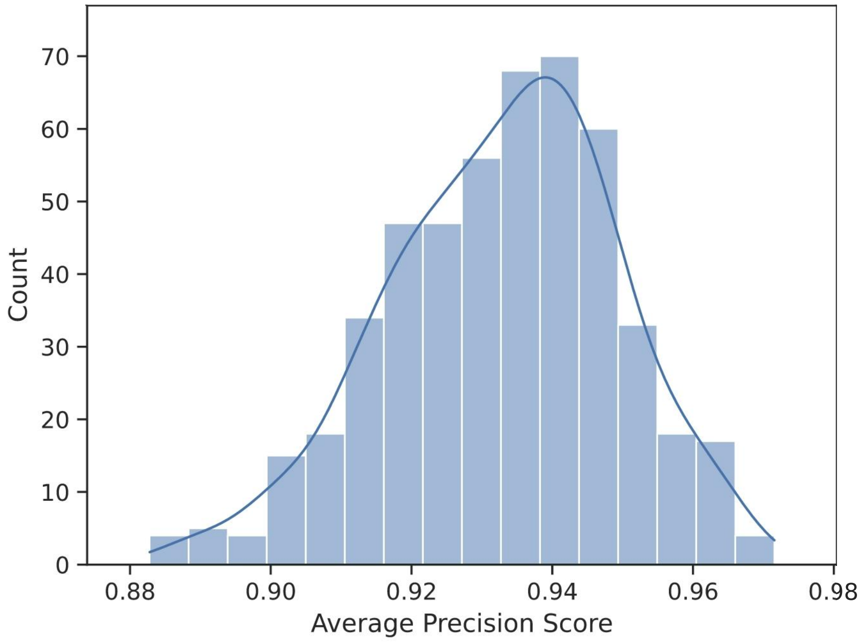


Figure 13 – Distribution of the test AUPRC for 500 refits of an XGBoost model with default hyperparameters following 500 80-20 train-test splits on the breast lesion dataset.

particular, he holds that data splitting hides the large instability associated with many feature selection algorithms.

A third approach lies in using resampling methods that simulate the acquisition of new data. The two most commonly discussed methods are cross-validation[2] [LM68; MT68; Sto74] and the bootstrap [Efr79; Efr83]. In this thesis we focus on cross-validation, however the bootstrap is equally applicable. In our context, cross-validation consists in partitioning the dataset into $k$ folds of roughly the same size and then, for each fold, performing the following: training a model on all folds except the one given and then testing the obtained model on the given fold. The cross-validated score is then the average of the scores obtained across the $k$ tests.

$k$-fold cross-validation can be used for model selection and validation in two ways. The first involves performing cross-validation for each model, and then choosing and reporting the metrics of the best performing model. The final model is then trained on the whole dataset. Alternatively, one might use a nested cross-validation approach, where model selection (and specially tuning) is seen as an integral part of model fitting. In this approach, an outer cross-validation loop is performed for model validation while an inner cross-validation is performed for selecting the model to be tested on each test fold. Similarly to simple cross-validation, when the best hyperparameter choice for a given fold is found, the model is refitted to the whole $(k-1)$ folds. Then, the refitted model is evaluated on the testing fold of the given iteration of the outer cross-validation. In the

[2] The reference to Tukey is accessible through his collected works [IMS02].

end, a simple cross-validation is performed on the full dataset and the reported scores are those from the simple cross-validation.

Cawley and Talbot [CT10] empirically motivate the usage of the second, more compute-intesive, approach to cross-validation as a way to test for overfitting to the model selection criterion. In particular, they demonstrate that one is able to perform hyperparameter tuning on kernel ridge regression to the point that test performance stops increasing, and instead decreases significantly, while the model selection criterion continues to improve. It is important to note that nested cross-validation doesn't solve the overfitting issue, it only tests for it. To actively take measure to prevent overfitting during model selection, one might either penalize for complexity or establish some early stopping criterion.

The standard recommendations for $k$-fold cross-validation are to use either $k = 5$ or $k = 10$ [HTF09; Koh95], and to perform multiple repetitions of the procedure by randomly shuffling the dataset into new folds [Har15]. It is commonly recommended that 100 repetitions be performed for 5-fold cross-validation and 50 repetitions be performed for 10-fold cross-validation [Har15].

It is important to note that train-test split and repeated $k$-fold cross-validation estimate different things [HTF09, Section 7.12]. Using the notation of the last chapter, the split estimates $R(\hat{h}_m)$, whereas cross-validation estimates $\mathbb{E}_{(x_1,y_1),\ldots,(x_m,y_m)\overset{iid}{\sim}\mathbb{P}}[R(\hat{h}_m)]$. In essence, the split validates a model, while cross-validation validates the entire model-building procedure. In particular, when validating a model that hasn't undergone hyperparameter tuning or feature selection, repeated $k$-fold cross-validation validates the learning algorithm itself. For instance, in our scenario, it serves to confirm the effectiveness of the XGBoost algorithm in using $(k-1)m/k$ training points to predict $m/k$ test points.

It should at last be mentioned that, in the absence of a mathematical model for the nature of the observed data, the previously outlined suggestions are to be interpreted as rules of thumb. Indeed, most of these recommendations stem from empirical experiments over a handful of specific datasets, and are not necessarily applicable to a particular problem found in practice. For a detailed discussion on what is empirical knowledge and what comes with a mathematical proof attached, see the comprehensive survey "A survey of cross-validation procedures for model selection" [AC10].